

Bayesian Analysis of Binary Prediction Tree Models for Retrospectively Sampled Outcomes

JENNIFER PITTMAN*, ERICH HUANG†, JOSEPH R NEVINS† & MIKE WEST*

January 2003

**Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA.*

†*Department of Molecular Genetics & Microbiology, Duke University, Durham, NC 27710, USA.*

SUMMARY

Classification tree models are flexible analysis tools which have the ability to evaluate interactions among predictors as well as generate predictions for responses of interest. We present a Bayesian approach to classification tree analysis in the specific context of a binary response Z with potentially very many candidate predictors x_i , and in which the data arise from a retrospective case-control design. This scenario is common in studies concerning gene expression data, which is a key motivating example context. The design issues are incorporated into the tree models via the use of underlying Dirichlet process priors on the distributions of predictor variables conditional on the response. This prior model influences the generation of trees through Bayes' factor based tests of association that determine significant binary partitions of nodes during a process of forward generation of trees. We describe this constructive process and discuss questions of generating and combining multiple trees via Bayesian model averaging for prediction. Additional discussion of parameter selection and sensitivity is given in the context of two examples, one of which concerns prediction of breast tumour status utilizing high-dimensional gene expression data; the examples demonstrate the exploratory/explanatory uses of such models as well as their primary utility in prediction.

Key Words: Bayesian analysis; binary classification tree; bioinformatics; retrospective sampling; case-control design; metagenes; molecular classification; predictive classification.

1 Introduction and Context

We discuss the generation and exploration of classification tree models, with particular interest in problems involving many predictors. The key motivating application is molecular phenotyping using gene expression and other forms of molecular data as predictors of a clinical or physiological state. We address the specific context of a binary response Z and many predictors x_i , and in which the data arise via a retrospective case-control design, i.e., observations are sampled retrospectively from a study where the numbers of 0/1 values in the response data are fixed by design. This is a very common context and has become particularly interesting in studies aiming to relate large-scale gene expression data (the predictors) to binary outcomes, such as a risk group or disease state (West et al., 2001). Breiman (2001) gives a useful discussion of recent developments in tree modeling and also an interesting gene expression example. Our focus here is on Bayesian analysis of this retrospective binary context.

Our analysis addresses and incorporates the retrospective case-control design issues in the assessment of association between predictors and outcome with nodes of a tree. With categorical or continuous covariates, this is based on an underlying non-parametric model for the conditional distribution of predictor values given outcomes, consistent with the retrospective case-control design. We use sequences of Bayes' factor based tests of association to rank and select predictors that define significant splits of nodes, and that provide an approach to forward generation of trees that is generally conservative in producing trees that are effectively self-pruning. We implement a tree-spawning method to generate multiple trees with the aim of finding classes of trees with high marginal likelihood, and prediction is based on model averaging, i.e., weighting predictions of trees by their implied posterior probabilities. Posterior and predictive distributions are evaluated at each node and the leaves of each tree, and feed into both the evaluation and interpretation tree by tree, and the averaging of predictions across trees for future cases to be predicted.

We give two examples, one concerning the prediction of levels of a binary representation of fat content of biscuits based on reflectance spectral measures of the raw dough (Brown et al., 1999; West, 2003). The second example concerns gene expression profiling using DNA microarray data as predictors of a clinical state in breast cancer. It is this latter problem area that motivated this work. The example of estrogen receptor (ER) status prediction given here demonstrates not only predictive value but also the utility of the tree modeling framework in aiding exploratory analysis that identifies multiple, related aspects of gene expression patterns related to a binary outcome, with some interesting interpretation and insights. This example also illustrates the use of what we term metagene factors – multiple, aggregate measures of complex gene expression patterns – in a predictive modeling context (West et al., 2001; Huang et al., 2003).

2 Model Context and Methodology

Data $\{Z_i, \mathbf{x}_i\}$ ($i = 1, \dots, n$) have been sampled retrospectively on a binary response variable Z and a p -dimensional covariate vector \mathbf{x} . The 0/1 response totals are fixed by design. Each predictor variable x_j could be binary, discrete or continuous.

2.1 Bayes' factor measures of association

At the heart of a classification tree is the assessment of association between each predictor and the response in subsamples, and we first consider this at a general level in the full sample. For any chosen

single predictor x , a specified threshold τ on the levels of x organizes the data into the 2×2 table

	$Z = 0$	$Z = 1$	
$x \leq \tau$	n_{00}	n_{01}	N_0
$x > \tau$	n_{10}	n_{11}	N_1
	M_0	M_1	

With column totals fixed by design, the categorized data is properly viewed as two Bernoulli sequences within the two columns, hence sampling densities

$$p(n_{0z}, n_{1z} | M_z, \theta_{z,\tau}) = \theta_{z,\tau}^{n_{0z}} (1 - \theta_{z,\tau})^{n_{1z}}$$

for each column $z = 0, 1$. Here, of course, $\theta_{0,\tau} = Pr(x \leq \tau | Z = 0)$ and $\theta_{1,\tau} = Pr(x \leq \tau | Z = 1)$. A test of association of the thresholded predictor with the response will be based on assessing the difference between these Bernoulli probabilities.

The natural Bayesian approach is via the Bayes' factor B_τ comparing the null hypothesis $\theta_{0,\tau} = \theta_{1,\tau}$ to the full alternative $\theta_{0,\tau} \neq \theta_{1,\tau}$. We adopt the standard conjugate beta prior model and require that the null hypothesis be nested within the alternative. Thus, assuming $\theta_{0,\tau} \neq \theta_{1,\tau}$, we take $\theta_{0,\tau}$ and $\theta_{1,\tau}$ to be independent with common prior $Be(a_\tau, b_\tau)$ with mean $m_\tau = a_\tau / (a_\tau + b_\tau)$. On the null hypothesis $\theta_{0,\tau} = \theta_{1,\tau}$, the common value has the same beta prior. The resulting Bayes' factor in favor of the alternative over the null hypothesis is then simply

$$B_\tau = \frac{\beta(n_{00} + a_\tau, n_{10} + b_\tau)\beta(n_{01} + a_\tau, n_{11} + b_\tau)}{\beta(N_0 + a_\tau, N_1 + b_\tau)\beta(a_\tau, b_\tau)}.$$

As a Bayes' factor, this is calibrated to a likelihood ratio scale. In contrast to more traditional significance tests and also likelihood ratio approaches, the Bayes' factor will tend to provide more conservative assessments of significance, consistent with the general conservative properties of proper Bayesian tests of null hypotheses (Selke et al., 2001).

In the context of comparing predictors, the Bayes' factor B_τ may be evaluated for all predictors and, for each predictor, for any specified range of thresholds. As the threshold varies for a given predictor taking a range of (discrete or continuous) values, the Bayes' factor maps out a function of τ and high values identify ranges of interest for thresholding that predictor. For a binary predictor, of course, the only relevant threshold to consider is $\tau = 0$.

2.2 Model consistency with respect to varying thresholds

A key question arises as to the consistency of this analysis as we vary the thresholds. By construction, each probability $\theta_{z,\tau}$ is a non-decreasing function of τ , a constraint that must be formally represented in the model. The key point is that the beta prior specification must formally reflect this. To see how this is achieved, note first that $\theta_{z,\tau}$ is in fact the cumulative distribution function of the predictor values x , conditional on $Z = z$, ($z = 0, 1$), evaluated at the point $x = \tau$. Hence the *sequence* of beta priors, $Be(a_\tau, b_\tau)$ as τ varies, represents a set of marginal prior distributions for the corresponding set of values of the cdfs. It is immediate that the natural embedding is in a non-parametric Dirichlet process model for the complete cdf. Thus the threshold-specific beta priors are consistent, and the resulting sets of Bayes' factors are comparable as τ varies, under a Dirichlet process prior with the betas as marginals. The required constraint is that the prior mean values m_τ are themselves values of a cumulative distribution function on the range of x , one that defines the prior mean of each θ_τ as a function. Thus, we simply rewrite the beta parameters (a_τ, b_τ) as $a_\tau = \alpha m_\tau$ and $b_\tau = \alpha(1 - m_\tau)$ for a specified prior mean value m_τ , where α is the prior precision (or "total mass") of the underlying Dirichlet process model. Note

that this specializes to a Dirichlet distribution when x is discrete on a finite set of values, including special cases of ordered categories (such as arise if x is truncated to a predefined set of bins), and also the extreme case of binary x when the Dirichlet is a simple beta distribution.

2.3 Generating a tree

The above development leads to a formal Bayes' factor measure of association that may be used in the generation of trees in a forward-selection process as implemented in traditional classification tree approaches. Consider a single tree and the data in a node that is a candidate for a binary split. Given the data in this node, construct a binary split based on a chosen (predictor, threshold) pair (x, τ) by (a) finding the (predictor, threshold) combination that maximizes the Bayes' factor for a split, and (b) splitting if the resulting Bayes' factor is sufficiently large. By reference to a posterior probability scale with respect to a notional 50:50 prior, Bayes' factors of 2.2, 2.9, 3.7 and 5.3 correspond, approximately, to probabilities of .9, .95, .99 and .995, respectively. This guides the choice of threshold, which may be specified as a single value for each level of the tree. We have utilized Bayes' factor thresholds of around 3 in a range of analyses, as exemplified below. Higher thresholds limit the growth of trees by ensuring a more stringent test for splits.

The Bayes' factor measure will always generate less extreme values than corresponding generalized likelihood ratio tests or significance testing (p -value) based approaches, and this can be especially marked when the sample sizes M_0 and M_1 are low. Thus the propensity to split nodes is always generally lower than with traditional testing methods, especially with lower sample sizes, and the approach tends to be more conservative in extending existing trees. Post-generation pruning is therefore generally much less of an issue, and can in fact generally be ignored.

Having generated a "current" tree, we run through each of the existing terminal nodes one at a time, and assess whether or not to create a further split at that node, stopping based on the above Bayes' factor criterion. Unless samples are very large (thousands) typical trees will rarely extend to more than three or four levels.

2.4 Inference and prediction with a single tree

Index the root node of any tree by zero, and consider the full data set of n observations, representing M_z outcomes with $Z = z$ in $0, 1$. Label successive nodes sequentially: splitting the root node, the left branch terminates at node 1, the right branch at node 2; splitting node 1, the consequent left branch terminates at node 3, the right branch at node 4, and so forth. Any node in the tree is labeled numerically according to its "parent" node; that is, a node j splits into two children, namely the (left, right) children $(2j + 1, 2j + 2)$. At level m of the tree ($m = 0, 1, \dots$) the candidate nodes are, from left to right, $2^m - 1, 2^m, \dots, 2^{m+1} - 2$.

Suppose we have generated a tree with m levels; the tree has some number of terminal nodes up to the maximum possible of $L = 2^{m+1} - 2$. Inference and prediction involve computations for *branch probabilities* and the predictive probabilities for new cases that these underlie. We detail this for a specific path down the tree, i.e., a sequence of nodes from the root node to a specified terminal node.

First, consider a node j that is split based on a (predictor, threshold) pair labeled (x_j, τ_j) (note that we use the node index to label the chosen predictor, for clarity). Extend the notation of Section 2.1 to include the subscript j indexing this node. Then the data at this node involve M_{0j} cases with $Z = 0$ and M_{1j} cases with $Z = 1$, and based on the chosen (predictor, threshold) pair (x_j, τ_j) , these samples split into cases $n_{00j}, n_{01j}, n_{10j}, n_{11j}$. The implied conditional probabilities $\theta_{z,\tau,j} = Pr(x_j \leq \tau_j | Z = z)$, for $z = 0, 1$, are the *branch probabilities* defined by such a split (note that these are also conditional on the tree and data subsample in this node, though the notation does not explicitly reflect this for clarity).

These are uncertain parameters and, following the development of Section 2.1, have specified beta priors, now also indexed by parent node j , i.e., $Be(a_{\tau,j}, b_{\tau,j})$. Assuming the node is split, the two sample Bernoulli setup implies conditional posterior distributions for these branch probability parameters: they are independent with posterior beta distributions

$$\theta_{0,\tau,j} \sim Be(a_{\tau,j} + n_{00j}, b_{\tau,j} + n_{10j}) \quad \text{and} \quad \theta_{1,\tau,j} \sim Be(a_{\tau,j} + n_{01j}, b_{\tau,j} + n_{11j}).$$

These distributions allow inference on branch probabilities, and feed into the predictive inference computations as follows.

Consider predicting the response Z^* of a new case based on the observed set of predictor values \mathbf{x}^* . The specified tree defines a unique path from the root to the terminal node for this new case. To predict requires that we compute the posterior predictive probability for $Z^* = 0/1$, which we do by following \mathbf{x}^* down the tree to the implied terminal node, and sequentially building up the relevant likelihood ratio defined by successive (predictor, threshold) pairs.

For example and specificity, suppose that the predictor profile of this new case is such that the implied path traverses nodes 0, 1, 4, and 9, terminating at node 9. This path is based on a (predictor, threshold) pair (x_0, τ_0) that defines the split of the root node, (x_1, τ_1) that defines the split of node 1, and (x_4, τ_4) that defines the split of node 4. The new case follows this path as a result of its predictor values, in sequence: $(x_0^* \leq \tau_0)$, $(x_1^* > \tau_1)$ and $(x_4^* \leq \tau_4)$. The implied likelihood ratio for $Z^* = 1$ relative to $Z^* = 0$ is then the product of the ratio of branch probabilities to this terminal node, namely

$$\lambda^* = \frac{\theta_{1,\tau_0,0}}{\theta_{0,\tau_0,0}} \times \frac{(1 - \theta_{1,\tau_1,1})}{(1 - \theta_{0,\tau_1,1})} \times \frac{\theta_{1,\tau_9,9}}{\theta_{0,\tau_9,9}}.$$

Hence, for any specified prior probability $Pr(Z^* = 1)$, this single tree model implies that, as a function of the branch probabilities, the updated probability π^* is, on the odds scale, given by

$$\frac{\pi^*}{(1 - \pi^*)} = \lambda^* \frac{Pr(Z^* = 1)}{Pr(Z^* = 0)}.$$

The retrospective case-control design provides no information about $Pr(Z^* = 1)$ so it is up to the user to specify this or examine a range of values; one useful summary is obtained by simply taking a 50:50 prior odds as benchmark, whereupon the posterior probability is

$$\pi^* = \lambda^* / (1 + \lambda^*).$$

Prediction follows by estimating π^* based on the sequence of conditionally independent posterior distributions for the branch probabilities that define it. For example, simply “plugging-in” the conditional posterior means of each θ . will lead to a plug-in estimate of λ^* and hence π^* . The full posterior for π^* is defined implicitly as it is a function of the θ . Since the branch probabilities follow beta posteriors, it is trivial to draw Monte Carlo samples of the θ . and then simply compute the corresponding values of λ^* and hence π^* to generate a posterior sample for summarization. This way, we can evaluate simulation-based posterior means and uncertainty intervals for π^* that represent predictions of the binary outcome for the new case.

2.5 Generating and weighting multiple trees

In considering potential (predictor, threshold) candidates at any node, there may be a number with high Bayes’ factors, so that multiple possible trees with different splits at this node are suggested. With continuous predictor variables, small variations in an “interesting” threshold will generally lead

to small changes in the Bayes’ factor – moving the threshold so that a single observation moves from one side of the threshold to the other, for example. This relates naturally to the need to consider thresholds as parameters to be inferred; for a given predictor x , multiple candidate splits with various different threshold values τ reflect the inherent uncertainty about τ , and indicate the need to generate multiple trees to adequately represent that uncertainty. Hence, in such a situation, the tree generation can spawn multiple copies of the “current” tree, and then each will split the current node based on a different threshold for this predictor. Similarly, multiple trees may be spawned this way with the modification that they may involve different predictors.

In problems with many predictors, this naturally leads to the generation of many trees, often with small changes from one to the next, and the consequent need for careful development of tree-managing software to represent the multiple trees. In addition, there is then a need to develop inference and prediction in the context of multiple trees generated this way. The use of “forests of trees” has recently been urged by Breiman (2001), and in references there, and our perspective endorses this. The rationale here is quite simple: node splits are based on specific choices of what we regard as parameters of the overall predictive tree model, the (predictor, threshold) pairs. Inference based on any single tree chooses specific values for these parameters, whereas statistical learning about relevant trees requires that we explore aspects of the posterior distribution for the parameters (together with the resulting branch probabilities).

Within the current framework, the forward generation process allows easily for the computation of the resulting relative likelihood values for trees, and hence to relevant weighting of trees in prediction. For a given tree, identify the subset of nodes that are split to create branches. The overall marginal likelihood function for the tree is the product of component marginal likelihoods, one component from each of these split nodes. Continue with the notation of Section 2.1 but, again, indexed by any chosen node j . Conditional on splitting the node at the defined (predictor, threshold) pair (x_j, τ_j) , the marginal likelihood component is

$$m_j = \int_0^1 \int_0^1 \prod_{z=0,1} p(n_{0zj}, n_{1zj} | M_{zj}, \theta_{z,\tau_j,j}) p(\theta_{z,\tau_j,j}) d\theta_{z,\tau_j,j}$$

where $p(\theta_{z,\tau_j,j})$ is the $Be(a_{\tau,j}, b_{\tau,j})$ prior for each $z = 0, 1$. This clearly reduces to

$$m_j = \prod_{z=0,1} \frac{\beta(n_{0zj} + a_{\tau,j}, n_{1zj} + b_{\tau,j})}{\beta(a_{\tau,j}, b_{\tau,j})}.$$

The overall marginal likelihood value is the product of these terms over all nodes j that define branches in the tree. This provides the relative likelihood values for all trees within the set of trees generated. As a first reference analysis, we may simply normalize these values to provide relative posterior probabilities over trees based on an assumed uniform prior. This provides a reference weighting that can be used to both assess trees and as posterior probabilities with which to weight and average predictions for future cases.

3 Example: Analysis of Biscuit Dough Data

A first example concerns biscuit dough data (Osborne et al., 1984; Brown et al., 1999; West, 2003) in which interest lies in relating aspects of near infrared (NIR) spectra of dough to the fat content of the resulting biscuits. The data set provides 78 samples, of which 39 are taken as training data and the remaining 39 as validation cases to be predicted, precisely as in Brown et al. (1999) and West (2003). We take the binary outcome to be 0/1 according to whether the measured fat content exceeds

a threshold; the threshold is chosen retrospectively to be the mean of the sample of fat values. As predictors, we take each \mathbf{x}_i to comprise 300 values of the spectrum of dough sample i , augmented by the set of singular factors (principal components) of the 78 sample spectra, so that $p = 378$, with singular factors indexed $301, \dots, 378$. With no prior information concerning each \mathbf{x}_i , the prior mean on the cdf of any \mathbf{x}_i , conditional on $Z = z$, across given thresholds is specified to take values from a uniform cdf on the range of x_i and the parameter of the Dirichlet process prior is set at $\alpha = 1/2$, corresponding to a Jeffrey’s prior on the complete cdf of each \mathbf{x}_i (Box & Tiao, 1992).

The analysis was developed repeatedly, exploring aspects of model fit and prediction of the validation sample as we vary a number of control parameters. The particular parameters of key interest are the Bayes’ factor thresholds that define splits, and controls on the number of such splits that may be made at any one node. Across ranges of these control parameters we find, in this example, that there is a good degree of robustness, and exemplify results based on values that, in this and a range of other examples, are representative. We fix the Bayes’ factor threshold at 3 on the log scale, and explore two-level trees allowing at most 10 splits of the root node and then at most 4 splits of each of nodes 1 and 2. This allows up to 160 trees, and this analysis generated 148.

Many of the trees identified had one or two of the predictors in common, and represent variation in the threshold values for those predictors. Figures 1-3 display some summaries. Figure 1 is one of the 148 trees, split at the root node by the spectral predictor labeled factor 92 (corresponding to a wavelength of 1566nm). Multiple wavelength values appear in the 148 trees, with values close to this appearing often, reflecting the underlying continuity of the spectra. The key second level predictor is factor 305, one of the principal component predictors. The data are scatter plotted on these two predictors in Figure 2 with corresponding levels of the predictor-specific thresholds from this tree marked. Slight changes in the value of either threshold would alter the predictions for several observations, leading to slight changes in the Bayes’ factor value. This sensitivity of Bayes’ factor values to threshold variation is reflected in trees which share predictors and in the variability among predictions from such trees.

The data appears also against the three predictors in this tree in Figure 3. Evidently there is substantial overlap in predictor space between the 0/1 outcomes, and cases close to the boundaries defined by any single tree are hard to accurately predict. Nevertheless, in terms of posterior predictive probabilities for the 39 validation samples, accuracy is good. Simply thresholding the predictive probabilities at 0.5 we find that 18 of 20 (90%) low fat (blue) cases are “correctly” predicted, as are 19 of 20 (95%) high fat (red) cases.

Predictive accuracy is high in this example, which we stress is one that has considerable overlap between predictor patterns among the two outcome groups. This is a positive example of the use of the predictive tree approach in a context where standard methods, such as logistic regression, would be less useful. We end with a note that the 50:50 split of the 78 samples into training and validation sets followed the previous authors as references. Curious about this, we reran the analysis 500 times, each time randomly splitting the data 50:50 into training and validation samples. Predictive accuracy, as measured above, was generally not as good as reported for the initial sample split, varying from a little below 50% to 100% across this set of 500 analyses. The average accuracy for low fat (blue) cases was 80%, and that for high fat (red) cases 76%.

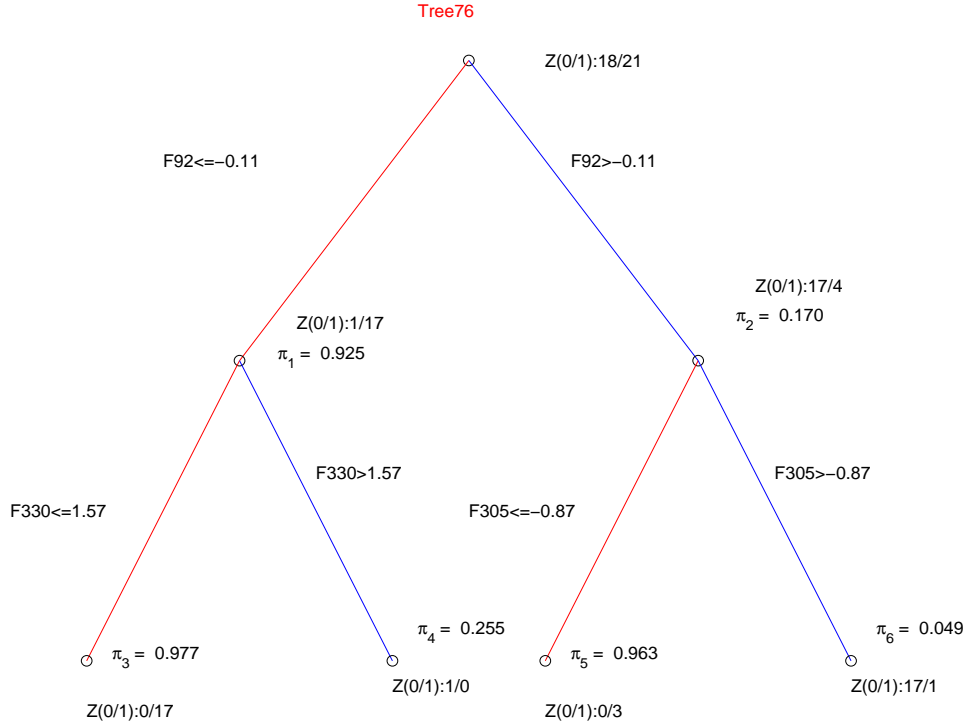


Figure 1: An example prediction tree for cookie fat outcomes.

The root node splits on predictor/factor 92, followed by two subsequent splits on additional predictors 330 and 305. The π values are point estimates of the predictive probabilities, π^* , of high fat versus low fat at each of the nodes, with suffixes simply indexing nodes. The labels $Z(0/1)$ indicate the numbers of low fat (0) and high fat (1) samples within each node, and the $F\#$ symbols indicate the thresholds that define the predictor based splits within each node.

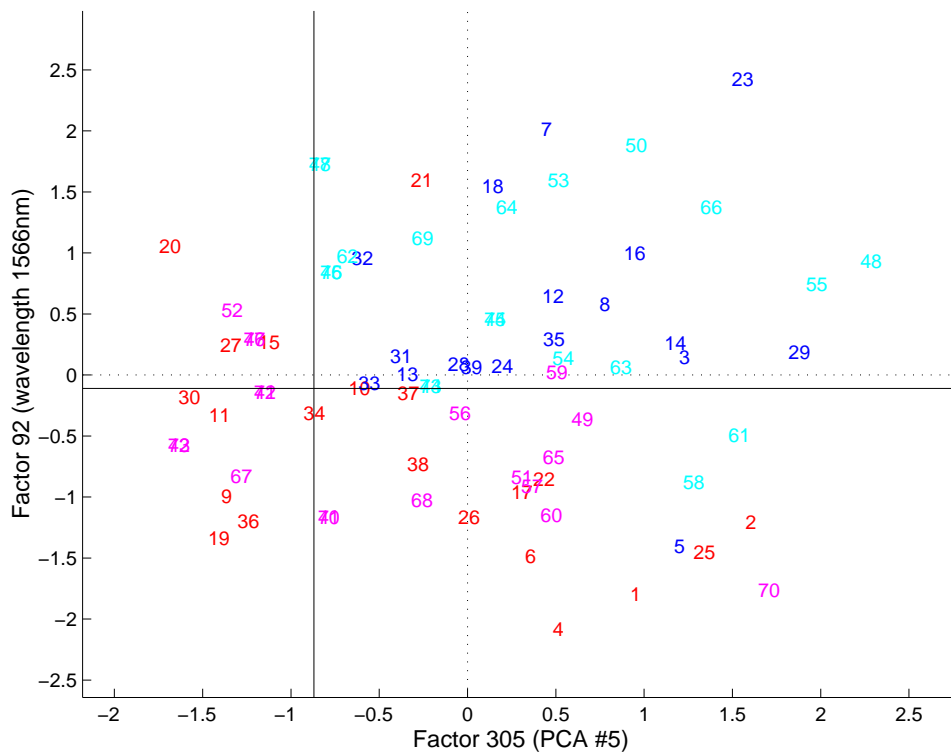


Figure 2: Two predictive factors in cookie dough analysis.

All samples are represented by index number in 1 – 78. Training data are denoted by blue (low fat) and red (high fat), and validation data by cyan (low fat) and magenta (high fat). The two full lines (black) demark the thresholds on the two predictors in this example tree.

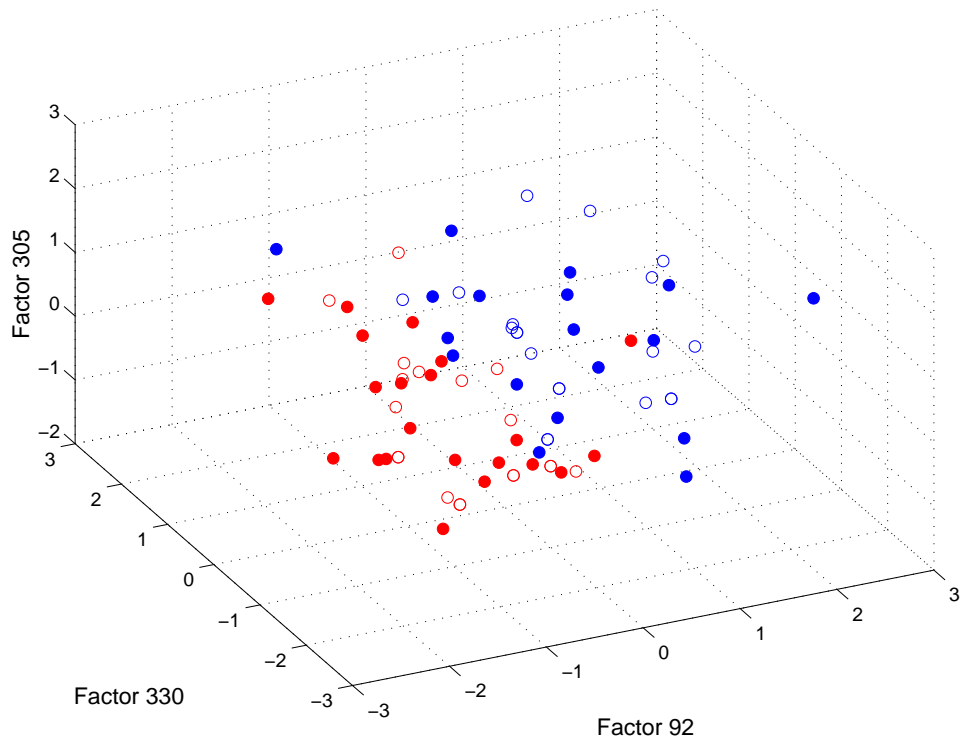


Figure 3: Scatter plot of cookie data on three factors in example tree.

Samples are denoted by blue (low fat) and red (high fat), with training data represented by filled circles and validation data by open circles.

4 Example: Metagene Expression Profiling

Our second example illustrates not only predictive utility but also exploratory use of the tree analysis framework in examining data structure. The context is primary breast cancer and the prediction of estrogen receptor (ER) status of breast tumors using gene expression data. West et al. (2001) presented an analysis of this data which involved binary regression, utilizing Bayesian generalized shrinkage approaches to factor regression (West, 2003); the model was a probit linear regression linking principal components of selected subsets of genes to the binary (ER positive/negative) outcomes.

We explore the same set of $n = 49$ samples here, using predictors based on *metagene* summaries of the expression levels of many genes. The evaluation and summarization of large-scale gene expression data in terms of lower dimensional factors of some form is being increasingly utilized for two main purposes: first, to reduce dimension from typically several thousand, or tens of thousands of genes; second, to identify multiple underlying “patterns” of variation across samples that small subsets of genes share, and that characterize the diversity of patterns evidenced in the full sample. Discussion of various factor model approaches appears in West (2003). In several recent studies we have used empirical metagenes, defined simply as principal components of clusters of genes; this is detailed in the Appendix, as it is of interest here only as it defines the predictor variables \mathbf{x} we utilize in the tree model example. It is, however, of much broader interest in gene expression profiling and related applications.

The data were sampled retrospectively and comprise 40 training samples and 9 validation cases. The training set was selected within a case-control framework to contain 20 ER positive samples and 20 ER negative samples. Among the validation cases, 3 were initial training samples that presented conflicting laboratory tests of the ER protein levels, so casting into question their actual ER status; these were therefore placed in the validation sample to be predicted, along with an initial 6 validation cases selected at random. These three cases are numbers 14, 31 and 33. The color coding in the graphs is based on the first laboratory test (immunohistochemistry). Additional samples of interest are cases 7, 8 and 11, cases for which the DNA microarray hybridizations were of poor quality, with the resulting data exhibiting major patterns of differences relative to the rest.

The metagene predictor has dimension $p = 491$. We generated trees based on a Bayes’ factor threshold of 3 on the log scale, allowing up to 10 splits of the root node and then up to 4 at each of nodes 1 and 2. The parameters of the Beta prior and the Dirichlet process prior were set as in the previous example. Some summaries appear in the following figures. Figures 4 and 5 display 3-D and pairwise 2-D scatterplots of three of the key metagenes, all clearly strongly related to the ER status and also correlated. There are in fact five or six metagenes that quite strongly associate with ER status and it is evident that they reflect multiple aspects of this major biological pathway in breast tumors. In our study reported in West et al. (2001), we utilized Bayesian probit regression models with singular factor predictors, and identified a single major factor predictive of ER. That analysis identified ER negative tumors 16, 40 and 43 as difficult to predict based on the gene expression factor model; the predictive probabilities of ER positive versus negative for these cases were near or above 0.5, with very high uncertainties reflecting real ambiguity.

What is very interesting in the current tree analysis, and particularly in relation to our prior regression analysis, is the identification of several metagene patterns that together combine to define an ER profile of tumors. When displayed as in Figures 4 and 5 these metagenes isolate these three cases as consistent with their designated ER negative status in some aspects, but conflicting and more consistent with the ER positive patterns on others. Metagene 347 is the dominant ER signature; the genes involved in defining this metagene include two representations of the ER gene, and several other genes that are coregulated with, or regulated by, the ER gene. Many of these genes appeared in the dominant factor in the regression prediction. This metagene is a strong discriminator of ER status, so it is no surprise that it shows up as defining root node splits in many high-likelihood trees. Metagene 347 also defines these

three cases – 16, 40 and 43 – as appropriately ER negative. However, a second ER associated metagene, number 352, defines a significant discrimination in which the three cases in question are much more consistent with ER positives. A number of genes, including the ER regulated PS2 protein and androgen receptors, play roles in this metagene, as they did in the factor regression; it is this second genomic pattern that, when combined together with the first as is implicit in the factor regression model, breeds conflicting information and results in ambivalent predictions with high uncertainty.

The tree model analysis here identifies multiple interacting patterns and allows easy access to displays such as these figures that provide insights into the interactions, and hence to interpretation of individual cases. In the full tree analysis, predictions based on averaging multiple trees are dominated by the root level splits on metagene 347, with all trees generated extending to two levels where additional metagenes define subsidiary branches. Due to the dominance of metagene 347, the three interesting cases noted above are perfectly in accord with ER negative status, and so are well predicted, even though they exhibit additional, subsidiary patterns of ER associated behavior identified in the figures. Figure 6 displays summary predictions in terms of point predictions of ER positive status with accompanying, approximate 90% intervals from the average of multiple tree models. The 9 validation cases are predicted based on the analysis of the full set of 40 training cases. The training cases are each predicted in an honest, cross-validation sense: each tumor is removed from the data set, the tree model is then refitted completely to the remaining 39 training cases only, and the hold-out case is predicted, i.e., treated as a validation sample. We note excellent predictive performance for both sets of samples. One ER negative, sample 31, is firmly predicted as having metagene expression patterns consistent with ER positive status; this is in fact one of the three cases for which the two laboratory tests conflicted. The other two such cases are number 33 and number 14, for which the predictions agree with the initial ER negative and ER positive test results, respectively. Case 8 is quite idiosyncratic, and the lack of conformity of expression patterns to ER status is almost surely due to major distortions in the DNA microarray data due to hybridization problems; the same issues arise with case 11, though case 7 is also a hybridization problem.

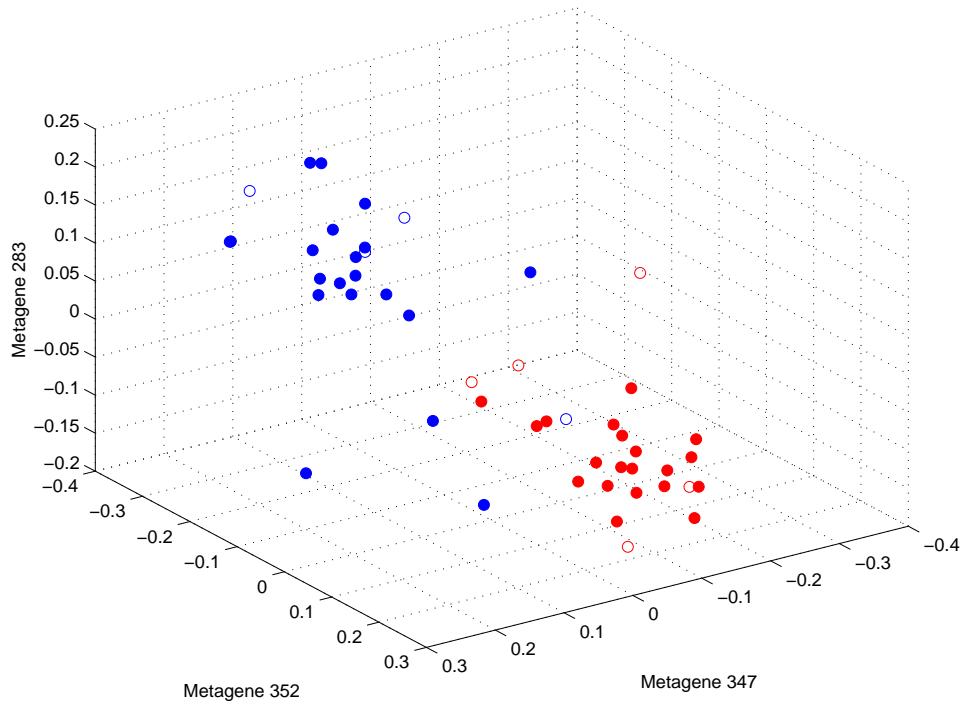


Figure 4: Three ER related metagenes in 49 primary breast tumors.

Samples are denoted by blue (ER negative) and red (ER positive), with training data represented by filled circles and validation data by open circles.

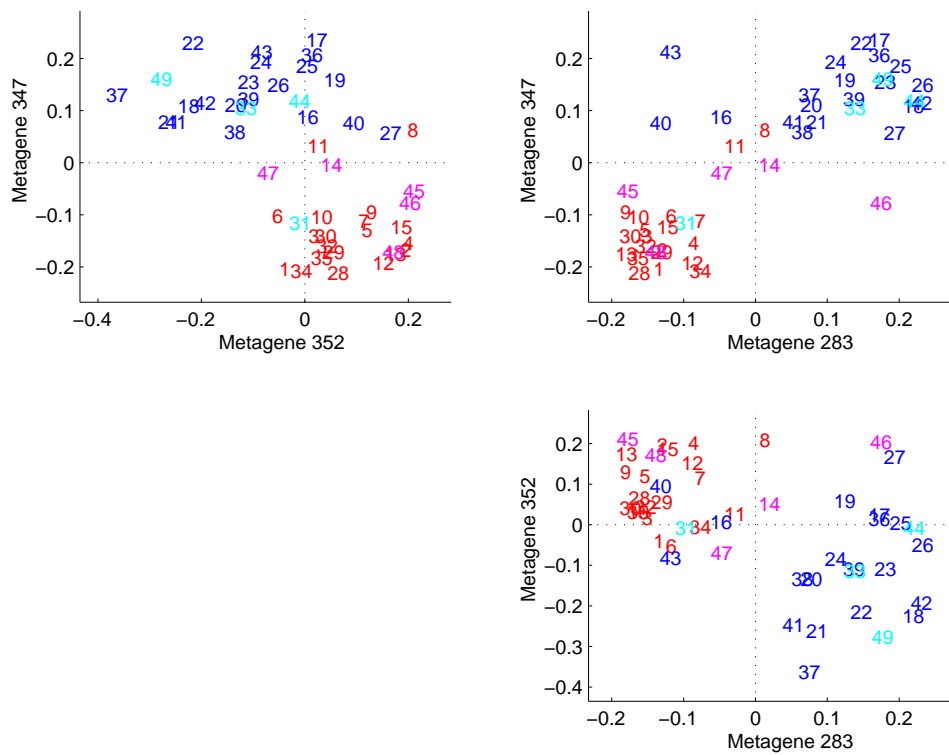


Figure 5: Three ER related metagenes in 49 primary breast tumors.

All samples are represented by index number in 1-78. Training data are denoted by blue (ER negative) and red (ER positive), and validation data by cyan (ER negative) and magenta (ER positive).

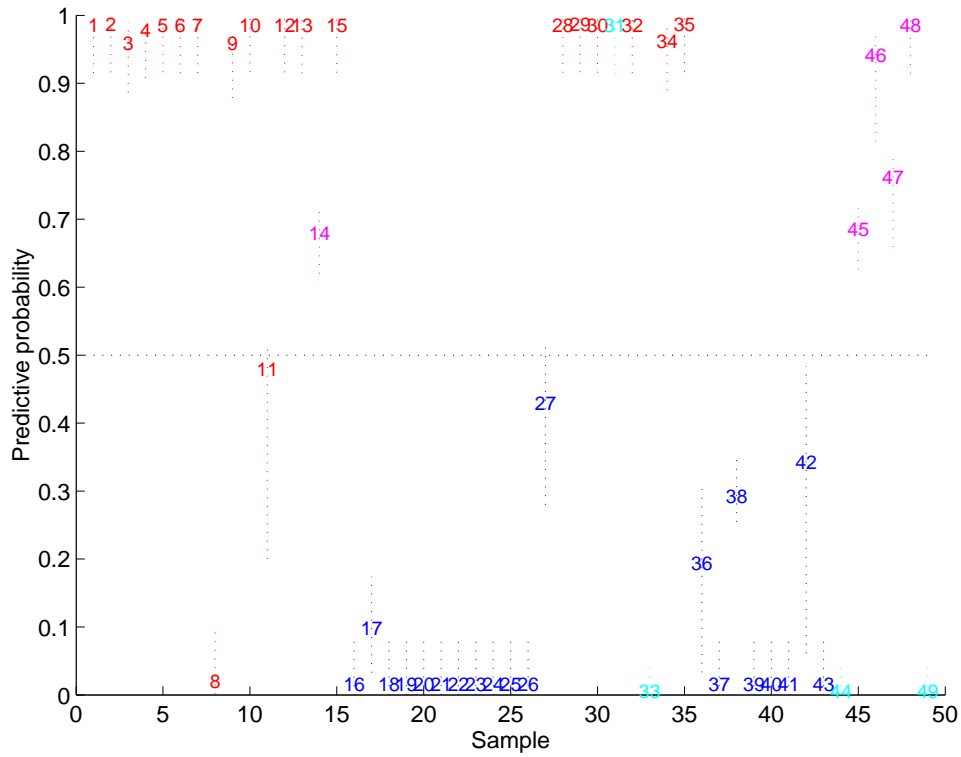


Figure 6: Honest predictions of ER status of breast tumors.

Predictive probabilities are indicated, for each tumor, by the index number on the vertical probability scale, together with an approximate 90% uncertainty interval about the estimated probability. All probabilities are referenced to a notional initial probability (incidence rate) of 0.5 for comparison. Training data are denoted by blue (ER negative) and red (ER positive), and validation data by cyan (ER negative) and magenta (ER positive).

5 Discussion

We have presented a Bayesian approach to classification tree analysis in the specific context of a binary response Z when the data arise via retrospective sampling. The sampling design is incorporated into the tree models by directly modelling the conditional distributions of predictor variables given the response, and defining a cascade of such distributions throughout successive nodes of any tree. In addition, we utilise nonparametric Dirichlet process priors for these conditional distributions; this leads to a flexible model for the distributions, while also ensuring consistency of model-based tests of association between outcomes and predictors that are thresholded. The resulting analysis provides a constructive Bayesian approach to predictive tree modelling.

The sensitivity of the Bayes' factor to (predictor, threshold) node split pair selection, i.e., to specific predictor choices and small changes in threshold values, is addressed by viewing splitting predictors and thresholds as parameters of a tree and capturing the variability in these parameters through tree-spawning and subsequent model averaging for inference and prediction. These methods are of particular importance in analyses involving many predictors, as is the case in studies involving gene expression data. We use the usual approach to tree generation that selects variables in a forward-selection process, growing trees from a null node. It is then natural to spawn multiple trees at a given node based on either the use of multiple candidate thresholds for a selected predictor variable as well as multiple candidate predictors. The resulting weighting and averaging over multiple trees then formally deals with these aspects of model uncertainty, albeit conditional on trees generated. We note that, though some progress has been made in developing stochastic simulation methods for Bayesian approaches to classification trees, the topic remains a very challenging research area, both conceptually and computationally, particularly in the context of more than a few predictors. Our interest lies in problems such as the molecular phenotyping example, where the numbers of predictors is very large. In such contexts, approaches based on the typical Bayesian MCMC format are simply infeasible and, we believe, will require a quite novel conceptual foundation before making them practicable. We are currently exploring the development of such ideas, and related approaches to stochastic search over tree space.

The examples highlight a number of methodological and substantive points, and demonstrate useful application in two retrospective (case-control) examples in the "large p , small n " paradigm. The tree models demonstrated strong predictive ability in both out-of-sample and one-at-a-time cross-validation contexts. This was achieved despite considerable overlap of outcome classes in the biscuit dough example and conflicting metagene information in the expression analysis example. The interaction of metagenes is useful not only for prediction but also for exploratory/explanatory purposes, e.g., suggesting possible reasons for ambiguous or uncertain predictions. The utility of the approach is further demonstrated in two recent application of these methods: some more directly clinical problems in breast cancer (Huang et al., 2003), and to gene discovery via molecular phenotyping in a cardiovascular disease context (Seo et al., 2003).

Appendix: Computing Metagene Expression Profiles

Metagenes are simple, summary measures of gene expression profiles derived as singular factors (principal components) of clusters of genes defined by standard clustering approaches. Assume a sample of n profiles of p genes. The specific construction used in the ER example here is detailed. The original data was developed on the early Affymetrix arrays with 7129 sequences, of which 7070 were used (following removal of Affymetrix controls from the data). The expression estimates used were log2 values of the signal intensity measures computed using the dChip software for post-processing Affymetrix output data; see Li & Wong (2001), and the software site <http://www.biostat.harvard.edu/complab/dchip/>.

We first screen genes to reduce the number by eliminating genes that show limited variation across samples or that are evidently expressed at low levels that are not detectable at the resolution of the gene expression technology used to measure levels. This removes noise and reduces the dimension of the predictor variable. Then, we used the k -means, correlated-based clustering as implemented in the xcluster software created by Gavin Sherlock (<http://genome-www.stanford.edu/sherlock/cluster.html>). We target a large number of clusters so as to capture multiple, correlated patterns of variation across samples, and generally small numbers of genes within clusters.

Following clustering, we extract the dominant singular factor (principal component) from each of the resulting clusters. Again, any standard statistical or numerical software package may be used for this; our analysis uses the efficient, reduced singular value decomposition function (*svd*) in the Matlab software environment (<http://www.mathworks.com/products/matlab>). In this example, with a target of 500 clusters, the xcluster software implementing the correlation-based k -means clustering produced $p = 491$ clusters. The corresponding p metagenes were then evaluated as the dominant singular factors of each of these clusters.

Acknowledgments

Research reported here was partially supported by NSF under grants DMS-0102227 and DMS-0112340, by SYNPAAC, NC, by the Koo Foundation Sun Yat-Sen Cancer Center Research Fund, and by the Keck Foundation through the Duke Keck Center for Neurooncogenomics. We are grateful to Ed Iversen of Duke University for useful discussions, and to Ming Liao for assistance with clustering software.

References

- BOX, G. & TIAO, G. (1992). *Bayesian Inference in Statistical Analysis*. New York: John Wiley and Sons.
- BREIMAN, L. (2001). Statistical modeling: The two cultures (with discussion). *Statist. Sci.* **16**, 199–225.
- BROWN, P., FEARN, T. & VANNUCCI, M. (1999). The choice of variables in multivariate regression: A non-conjugate bayesian decision theory approach. *Biometrika* **86**, 635–648.
- HUANG, E., CHENG, S. H., DRESSMAN, H., PITTMAN, J., TSOU, M. H., HORNG, C. F., BILD, A., IVERSEN, E. S., LIAO, M., CHEN, C. M., WEST, M., NEVINS, J. R. & HUANG, A. T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* .
- LI, C. & WONG, W. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98**, 31–36.
- OSBORNE, B., FEARN, T., MILLER, A. & DOUGLAS, S. (1984). Applications of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.* **35**, 99–105.
- SELKE, T., BAYARRI, M. & BERGER, J. (2001). Calibration of p -values for testing precise null hypotheses. *The American Statistician* **55**, 62–71.
- SEO, D. M., DRESSMAN, H., HERDERICK, E. E., IVERSEN, E. S., DONG, C., VATA, K., MILANO, C. A., NEVINS, J. R., PITTMAN, J., WEST, M. & GOLDSCHMIDT-CLERMONT, P. J. (2003). Gene expression phenotypes of atherosclerosis. Tech. rep., Institute of Statistics & Decision Sciences, and Computational & Applied Genomics Program, Duke University. Available at: www.cagp.duke.edu.
- WEST, M. (2003). Bayesian factor regression models in the “large p , small n ” paradigm. In *Bayesian Statistics 7*, J. M. Bernardo, M. J. Bayarri, J. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West, eds. Oxford University Press.
- WEST, M., BLANCHETTE, C., DRESSMAN, H., ISHIDA, S., SPANG, R., ZUZAN, H., MARKS, J. R. & NEVINS, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.* **98**, 11462–11467.