Biostatistics (2004), **5**, 4, *pp.* 587–601
doi: 10.1093/biostatistics/kxh011

Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes

JENNIFER PITTMAN[†]
Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708-0251, USA
jennifer@stat.duke.edu

ERICH HUANG
Department of Surgery, Duke University, Durham, NC 27708-0251, USA

JOSEPH NEVINS
Department of Molecular Genetics & Microbiology, Duke University, Durham, NC 27708-0251, USA

QUANLI WANG
Computational & Applied Genomics Program, Duke University, Durham, NC 27708-0251, USA

MIKE WEST
Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708-0251, 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 describe Bayesian analysis of a specific class of tree models in which binary response data arise from a retrospective case-control design. We are also particularly interested in problems with potentially very many candidate predictors. This scenario is common in studies concerning gene expression data, which is a key motivating example context. Innovations here include the introduction of tree models that explicitly address and incorporate the retrospective design, and the use of nonparametric Bayesian models involving Dirichlet process priors on the distributions of predictor variables. The model specification 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 an example which concerns prediction of breast tumour status utilizing high-dimensional gene expression data; the example demonstrates the exploratory/explanatory uses of such models as well as their primary utility in prediction. Shortcomings of the approach and comparison with alternative tree modelling algorithms are also discussed, as are issues of modelling and computational extensions.

Keywords: Bayesian analysis; Binary classification tree; Bioinformatics; Case-control design; Metagenes; Molecular classification; Predictive classification; Retrospective sampling; Tree models.

[†] To whom correspondence should be addressed.

1. INTRODUCTION

We discuss the generation and exploration of classification tree models, with particular interest in problems involving many predictors. One 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 to binary outcomes, such as a risk group or disease state (West *et al.*, 2001). Breiman (2001b) gives a useful discussion of recent developments in tree modelling 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 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.

Existing Bayesian approaches to tree modelling utilizing stochastic search methods have proven to be effective in real data examples (Chipman *et al.*, 1998; Denison *et al.*, 1998). These current approaches utilize ideas from MCMC though define effective stochastic search algorithms for plausible models rather than provably convergent MCMC methods for posterior sampling. Indeed, development of MCMC for full posterior sampling in tree models remains an open and very challenging problem even in contexts with very small numbers of candidate predictor variables. These existing methods scale very poorly, and so are ill-suited to problems of high-dimensional predictor spaces. In microarray applications it is not unusual to have thousands of potential predictors and the implementation of a simulation-based approach in such a context requires research advances in statistical computation. Our approach, in contrast, utilizes deterministic search that aims to efficiently generate many candidate tree models, and models of high likelihood. Related to classical approaches to tree model generation (Breiman *et al.*, 1984, Clark and Pregibon, 1992), such methods have previously been employed for gene expression analysis (Hastie *et al.*, 2001; Segal *et al.*, 2003; Boulesteix *et al.*, 2003) as both exploratory and predictive tools. Beyond the development of Bayesian analysis, with its direct focus on prediction and particularly on the evaluation of uncertainties in prediction via model averaging (Raftery *et al.*, 1997) over multiple trees, we present an approach and non-parametric model class that is specifically tailored to the retrospective sampling paradigm. We also highlight the exploratory uses of tree modelling in evaluation and exploration of predictors appearing across multiple, plausible models.

Following discussion and model development, we give an example concerning gene expression profiling using DNA microarray data as predictors of a clinical state in breast cancer. The example of estrogen receptor (ER) status prediction given here demonstrates not only predictive value but also the utility of the tree modelling 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 modelling context (West *et al.*, 2001; Huang *et al.*, 2003). We compare our results to those provided by random forests (Breiman, 2001a), as implemented in the widely available statistical software package R (Breiman *et al.*, 2004; R Development Core Team, 2003). Software implementing our tree modelling approach is available for download at (`http://www.cagp.duke.edu/~`).

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_i 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

	$Z = 0$	$Z = 1$	
$x \leqslant \tau$	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 are properly viewed as two Bernoulli sequences within the two columns, hence sampling densities

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

for each column $z = 0, 1$. Here, of course, $\theta_{0,\tau} = \Pr(x \leqslant \tau | Z = 0)$ and $\theta_{1,\tau} = \Pr(x \leqslant \tau | Z = 1)$. 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 $\theta_{0,\tau} \neq \theta_{1,\tau}$. The Bayes' factor is defined as the posterior odds of either the null or the alternative hypothesis when the prior probabilities of the two hypotheses are equal (Kass and Raftery, 1995). It can be viewed as a measure of the evidence provided for or against either hypothesis. In calculating the Bayes' factor 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 $\mathrm{Be}(a_\tau, b_\tau)$ 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 favour of the alternative over the null hypothesis is then

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

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. In the following we consider predictors which can be ordered. By construction, each probability $\theta_{z,\tau}$ is a non-decreasing function of τ and the key point is that the beta prior specification must formally reflect this constraint. 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 $x = \tau$. Typically we select the threshold values for a given predictor to be quantiles of the observed data values for the predictor. Hence the *sequence* of beta priors, $\mathrm{Be}(a_\tau, b_\tau)$ 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. The tree models that we consider each represent a recursive binary partition of the feature space into a set of rectangles (Hastie *et al.*, 2001). Initially the space is split into two regions, represented by nodes of the tree, where the variable and split-point (or threshold) are chosen to achieve the best fit. This splitting process is continued recursively on the tree nodes, resulting in a partition of the space into smaller subspaces represented by the leaves or terminal nodes of the tree.

Consider a single tree and data in a node that are 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 0.9, 0.95, 0.99 and 0.995, 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. Unless samples are very large (thousands) typical trees will rarely extend to more than three or four levels.

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. Due to the stepwise nature of the model search it is possible that good models could be missed, i.e. since the algorithm stops splitting a node as soon as the Bayes’ factor criterion is not met it may overlook subsequent splits that may lead to promising trees. In problems with a very large number of predictors it may be possible to vary the Bayes’ factor criterion across levels or increase the number of trees to aid in the search for high likelihood models.

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 labelled 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 candidates 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 labelled (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 $\theta_{z,\tau,j} = \Pr(x_j \leqslant \tau_j | Z = z)$, 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. $\mathrm{Be}(a_{\tau,j}, b_{\tau,j})$. These beta priors are indexed by the parent node because their values depend on the split variable and threshold at that node (not on the particular level or location of the node). 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

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

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

The use of independent priors is an approximation as we would expect dependence among predictors in certain scenarios, e.g. some genes may be known to perform similar functions or participate in similar biological pathways. As our interests lean toward a more automatic approach and for the sake of computational efficiency we have chosen to adopt independent priors and shift the subjective component of the method to the choice of predictors and the predictor space.

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, 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: $(x_0^* \leqslant \tau_0)$, $(x_1^* > \tau_1)$ and $(x_4^* \leqslant \tau_4)$. 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

$$\lambda^* = \frac{\theta_{1,\tau_0,0}}{\theta_{0,\tau_0,0}} \times \frac{(1-\theta_{1,\tau_1,1})}{(1-\theta_{0,\tau_1,1})} \times \frac{\theta_{1,\tau_4,4}}{\theta_{0,\tau_4,4}}.$$

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 taking a 50:50 prior odds as benchmark, whereupon the posterior probability is

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

In a case-control context if a new case were selected at random from the population, a useful estimate of $\Pr(Z^* = 1)$ could be determined from the prevalence of the disease (obtained from disease registries, for example). The prior odds may then be other than 50:50 and the expression for π^* would be obtained by replacing λ^* with $\lambda^*\Pr(Z^* = 1)/\Pr(Z^* = 0)$. However, in a cross-validation context similar to that of our example, the 50:50 choice seems reasonable.

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 $\theta_{\cdot}$ 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 $\theta_{\cdot}$. Since the branch probabilities follow beta posteriors, it is trivial to draw Monte Carlo samples of the $\theta_{\cdot}$ 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 our approach the maximum number of significant splits at each node that are carried through to children is limited by computational considerations, although this limit is high enough to allow for the construction of thousands of 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 (2001b), 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

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

where $p(\theta_{z,\tau_j,j})$ is the $\mathrm{Be}(a_{\tau,j}, b_{\tau,j})$ 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. Trees with more nodes will have lower marginal likelihood values unless the splits generating the additional nodes lead to a substantially improved model, providing an implicit penalty against many nodes. As a first reference analysis, we may simply normalize the likelihood 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: metagene expression profiling

Our 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 tumours 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 (Huang *et al.*, 2003; Seo *et al.*, 2003); 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 nine 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, three 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 six validation cases selected at random. These three cases are numbers 14, 31 and 33. If the model demonstrates the ability to predict the status of the six randomly selected samples then perhaps the model predictions for the three questionable cases can help eulcidate their true ER status. The colour 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. For comparison we modelled the data using the random forest package available in R (Breiman *et al.*, 2004; R Development Core Team, 2003) where the parameters (e.g. number of trees, minimum leaf size) were chosen to best match those in the Bayesian tree approach.

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. In other words, at the root node the algorithm will search through all candidate predictor/threshold combinations and create trees based on the most significant splits until 10 trees have been created or all combinations have been searched. This same process will be repeated at node 1 of each resulting tree, yielding up to four different splits of this node for each tree (a total of up to forty trees). Starting with these trees, the same process is again repeated at node 2, yielding up to four different splits of this node for each tree (a total of up to 160 trees). 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 and Tiao, 1992). The analysis was developed repeatedly, exploring aspects of model fit and prediction of the validation sample as we varied 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. By varying the Bayes' factor threshold between 1.0 and 4.0 and the number of splits from 1 to 50 at the root node and 1 to 5 at nodes 1 and 2 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.

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 and 2 display 3D and pairwise 2D 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 tumours. 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 tumours 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.

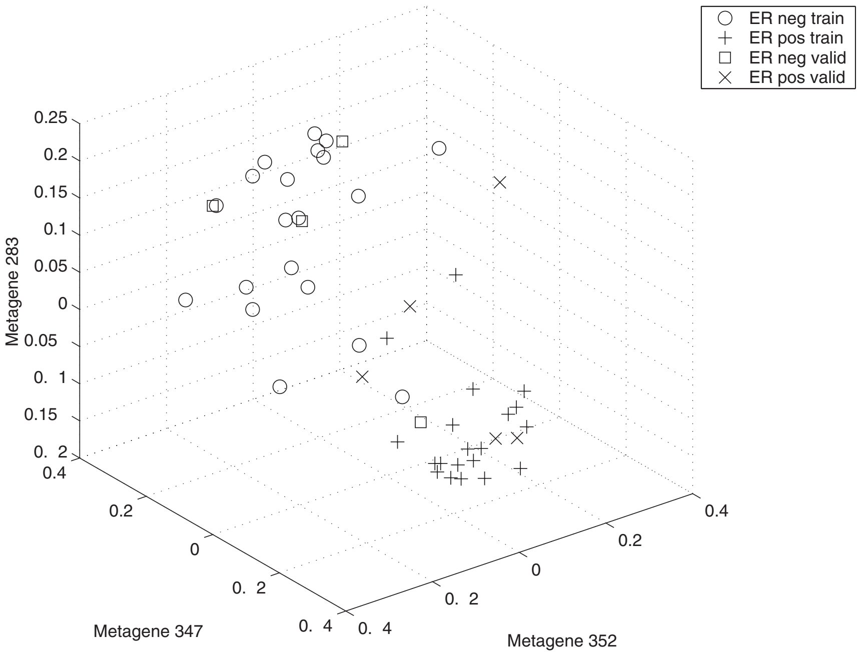

Fig. 1. Three ER related metagenes in 49 primary breast tumours. Samples are denoted by open symbols (ER negative) and closed symbols (ER positive), with training data represented by circles and plus signs and validation data by open squares and crosses.

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 tumours. When displayed as in Figures 1 and 2 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 as seen in the summary of the trees involved in the prediction of the validation cases (Figure 4); 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 and this metagene was also selected as the dominant predictor in the random forest trees. The random forest trees were run with parameter settings which allowed a comparable number of trees and minimum node size as the Bayesian tree implementation; the number of variables considered at each node was varied and the results were insensitive to values above 200. Metagene 347 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, which appears in trees from the cross-validation runs but not in the trees for the validation cases, 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 random forest trees does not identify metagene 352 (or any variable highly correlated with metagene 352)

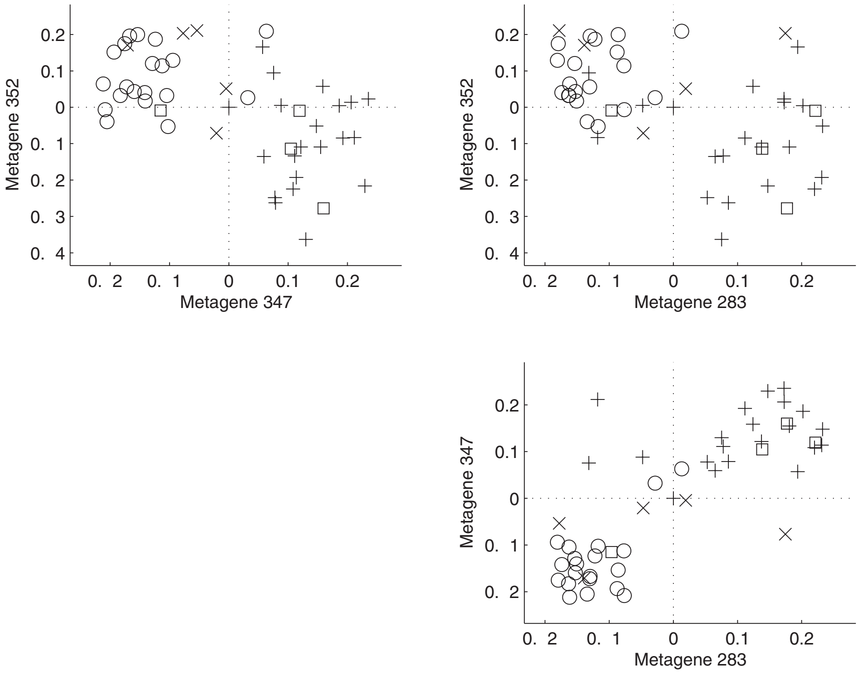

Fig. 2. Three ER related metagenes in 49 primary breast tumours. All samples are represented by index number in 1–49. Training data are denoted by circles (ER negative) and plus signs (ER positive), and validation data by squares (ER negative) and crosses (ER positive).

as a significant model predictor; instead they identified metagenes 283 and 402 that positively correlate with metagene 347. Note also that for the validation cases all of the trees received similar weights and hence made similar contributions to the model predictions. In this scenario it is important that our method properly account for model uncertainty, in the Bayesian framework. Although the random forest approach does provide a measure of variable importance it cannot account properly for model uncertainty within this framework.

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 behaviour identified in the figures. Figure 3 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 nine 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 tumour 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. For the random forests the training data predictions were based on the bootstrap samples while the validation cases were predicted based on the bootstrap trees based on all of the training cases. We note excellent predictive performance for the Bayesian trees on both sets of samples. One ER negative,

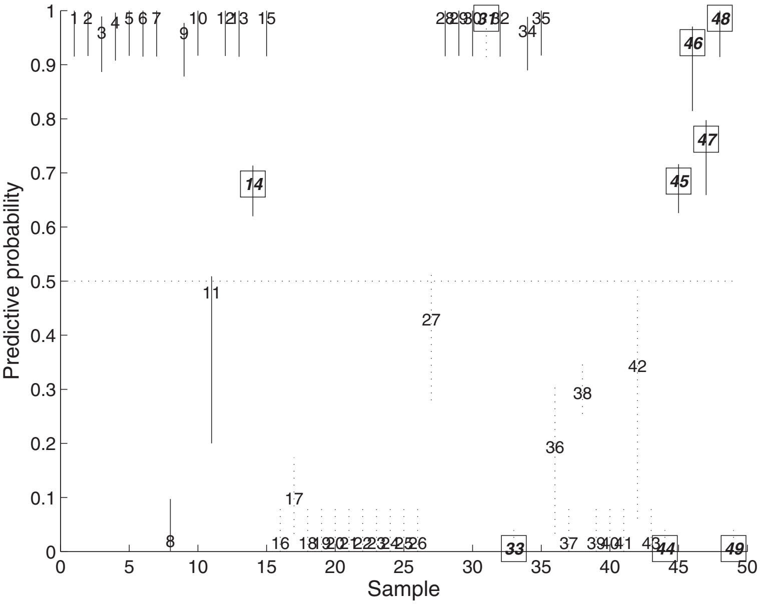

Fig. 3. Honest Bayesian tree predictions of ER status of breast tumours. Predictive probabilities are indicated, for each tumour, by the index number on the vertical probability scale, together with an approximate 90% uncertainty interval about the estimated probability. All probabilities are referenced to a notional initial probability (incidence rate) of 0.5 for comparison. Training data are denoted by light text and validation data by bold boxed text; ER negative samples have dotted uncertainty lines and ER positive samples have solid uncertainty lines.

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. The random forest results were similar to those of our Bayesian approach, on average, although uncertainty intervals for the predictions were not provided. 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.

The validation predictions are encouraging evidence that our method does not overfit the data. As further support of this, an experiment was conducted in which the rows and columns of the metagene expression matrix were randomly permuted. The tree models were fit to the permuted training samples and predictions were made for the validation cases. This process was repeated 100 times and the average prediction accuracy across runs was 55.56% (with 5% and 95% accuracy bounds of 22.22% and 77.78%, respectively). This result further demonstrates that with respect to out-of-sample prediction our method does not tend towards overfitting in this example.

4. 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,

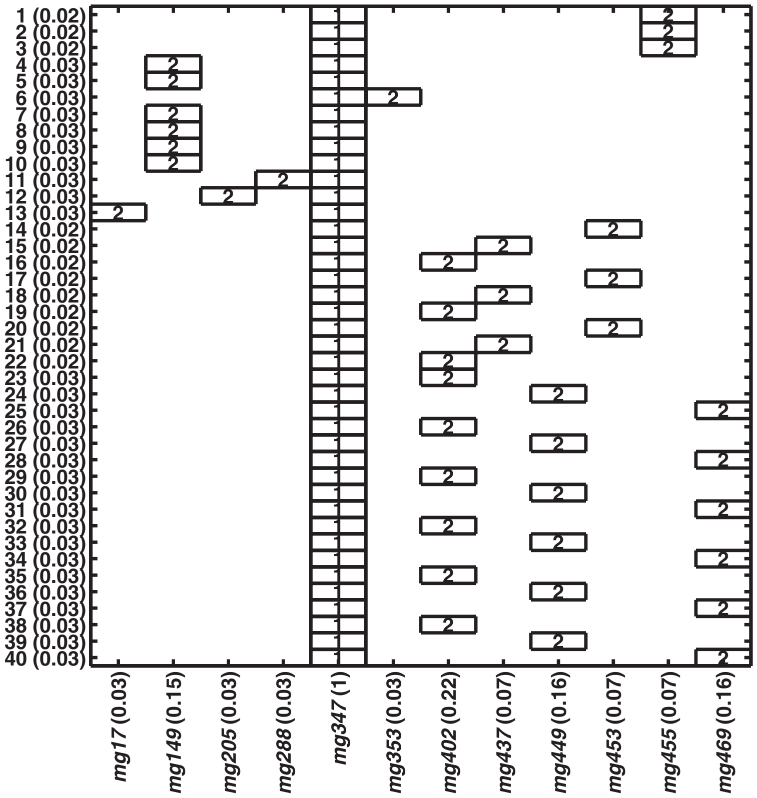

Fig. 4. Predictor variables in Bayesian genomic tree models for validation predictions of ER status of breast tumours. A summary of the level of the tree in which each variable appears and defines a node split. The numbers on the left simply index trees, and the probabilities in parentheses on the left indicate the relative weights of trees based on fit to the data. The probabilities associated with the predictor variables in parentheses on the horizontal scale are sums of the probabilities of trees in which each occurs, and so define overall weights indicating the relative importance of each variable to the overall model fit and consequent predictions.

and defining a cascade of such distributions throughout successive nodes of any tree. In addition, we utilize 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 or 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 issue of 'dilution' of the prior for a predictor with many splits is relevant to any tree approach. Since we do not place a prior weight on each predictor our method does not suffer from 'dilution' in this sense; actually, the issue is the reverse as predictors with more thresholds have a higher probability of being selected to define a tree split. In our example we selected the predictor thresholds as quantiles of the observed predictor values so each predictor has the same number of thresholds, hence avoiding this issue. However, it is a general issue for tree methods and an interesting area for future research.

It is possible that a particularly strong predictor with many candidate thresholds could dominate the tree search. We partially aleviate this by using the same number of thresholds for each predictor but if a particular predictor is dominant it may be useful to select subsets of predictors for modelling, as in random forests. This could lead to the involvement of more predictors in the model process but may obscure interactions between predictors and disallow the possibility of properly accounting for model uncertainty across predictors.

The example highlights a number of methodological and substantive points, and demonstrates useful application in a retrospective (case-control) example 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 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. Although the random forest implementation did provide reasonable predictions it could not properly account for model uncertainty nor match the utility of the Bayesian trees as an exploratory/explanatory tool. The utility of the approach described here is further demonstrated in two recent applications of these methods: 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).

Acknowledgments

Research reported here was partially supported by NSF under grants DMS-0102227 and DMS-0112340.

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 were 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 and 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 cluster, 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 cluster.

REFERENCES

BOULESTEIX, A.-L., TUTZ, G. AND STRIMMER, K. (2003). A CART approach to discover emerging patterns in microarray data. *Bioinformatics* **19**, 2465–2472.

BOX, G. AND TIAO, G. (1992). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

BREIMAN, L. (2001a). Random forests. *Machine Learning* **45**, 5–32.

BREIMAN, L. (2001b). Statistical modeling: The two cultures (with discussion). *Statistical Science* **16**, 199–225.

BREIMAN, L., CUTLER, A., LIAW, A. AND WIENER, M. (2004). *The randomForest Package*, version 4.0.7.. Vienna: R Foundation for Statistical Computing.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. AND STONE, C. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

CHIPMAN, H., GEORGE, E. AND MCCULLOCH, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association* **93**, 935–960.

CLARK, L. AND PREGIBON, D. (1992). Tree based models. Chambers, J. and Hastie, T. (eds), *Statistical Models in S*. Wadsworth: Pacific Grove, CA, pp. 377–420.

DENISON, D., MALLICK, B. AND SMITH, A. (1998). A Bayesian CART algorithm. *Biometrika* **85**, 363–377.

HASTIE, R., TIBSHIRANI, T. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

HASTIE, R., TIBSHIRANI, T., BOTSTEIN, D. AND BROWN, P. (2001). Supervised harvesting of expression trees. *Genome Biology* **2**, 1–12.

HUANG, E., CHENG, S. H., DRESSMAN, H., PITTMAN, J., TSOU, M. H., HORNG, C. F., BILD, A., IVERSEN, E. S., LIAO, M. AND CHEN, C. M., *et al.* (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596.

KASS, R. AND RAFTERY, A. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

LI, C. AND WONG, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* **98**, 31–36.

R DEVELOPMENT CORE TEAM (2003). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

RAFTERY, A., MADIGAN, D. AND HOETING, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.

SEGAL, E., SHAPIRA, M., REGEV, A., PE'ER, D., BOTSTEIN, D. AND KOLLER, D. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**, 166–176.

SELKE, T., BAYARRI, M. J. AND BERGER, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician* **55**, 62–71.

SEO, D., DRESSMAN, H., HERDERICK, E. E., IVERSEN, E. S., DONG, C., VATA, K., MILANO, C. A., NEVINS, J. R., PITTMAN, J., WEST, M. AND GOLDSCHMIDT-CLERMONT, P. J. (2003). Gene expression phenotypes of atherosclerosis. *Technical Report*, Available at: www.cagp.duke.edu. Durhan, NC: Institute of Statistics & Decision Sciences, and Computational & Applied Genomics Program, Duke University.

WEST, M. (2003). Bayesian factor regression models in the 'large p, small n' paradigm. Bernardo, J. M., Bayarri, M. J., Berger, J.O., Dawid, A. P., Heckerman, D., Smith, A. F. M. and West, M. (eds), *Bayesian Statistics 7*. Oxford: Oxford University Press, pp. 723–732.

WEST, M., BLANCHETTE, C., DRESSMAN, H., ISHIDA, S., SPANG, R., ZUZAN, H., MARKS, J. R. AND NEVINS, J. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* **98**, 11462–11467.

[*Received December 16, 2003; revised April 14, 2004; accepted for publication April 19, 2004*]