

Prediction Tree Models in Clinico-Genomics

Jennifer Pittman, Erich Huang, Joseph Nevins and Mike West
Duke University, ISDS
Durham, NC 27708, USA

Contact: mw@stat.duke.edu

1. Introduction

Classification tree models have ability to discover and evaluate interactions of multiple predictor variables, and define flexible, nonlinear predictive tools. We have developed tree models for clinical prediction studies with very high-dimensional gene expression data as candidate predictors. A first context is Bayesian tree models for predicting binary outcomes (as an example), that respects a retrospective (case-control) sampling design common in gene expression studies. A second context is survival modelling for problems such as disease recurrence. Key issues are approaches to tree construction, multiplicities, sensitivity of tree predictions, and the need to average predictions over multiple candidate models. Some of our disease studies use metagene predictors – aggregate gene expression signatures from clusters of genes – with clinical variables. We stress the utility of such tree models for gene and metagene data exploration, and the resulting identification of genes plausibly associated with clinical endpoints, as well as for clinico-genomic prediction.

2. Predictive Tree Modelling and Gene Expression Data as Predictors

A clinical outcome y is related to many predictors in a p -vector x via models $p(y|x)$. With genomic data, x includes expression levels of many genes and also *metagene* signatures – principal components derived from multiple clusters of genes (Huang et al, 2003, Pittman et al 2002), as well as non-genomic predictors. Tree models define specific distributions within nodes of multiple trees, defined by recursively splitting the data within each node according to a threshold on a chosen predictor, i.e., $\{x_i < \tau_i\}$ for some $i = 1, \dots, p$ (Breiman 2001 and references; Chipman et al 1998). Forward selection of tree models chooses node splits progressively “down” a tree based on optimising an association measure over $\{i, \tau_i\}$, and testing whether or not to split based on an assessment of significance of each split. Our association testing uses probability models, computing (conservative) Bayes’ factors to test a null hypothesis of a common distribution within a node relative to a split into two subpopulations. The form of the assessment depends heavily on context, as described below. A given tree delivers predictions at terminal nodes, and we weight across multiple trees by normalized tree likelihood functions. Multiple trees can be “spawned” at any node based on multiple choices of $\{i, \tau_i\}$, and this generates classes of trees for likelihood evaluations and combinations in prediction. Full details and example are given in Huang et al (2002), Pittman et al (2002), Pittman et al (2003).

3. Trees for Retrospective Sampling: Binary Outcomes Example

If data arises from a case-control study, we see samples from $p(x|y)$ for design-specified y values. In the example of binary y , at each node the assessment of association of y with a given predictor/threshold pair $\{x_i, \tau_i\}$, is then based on assessing differences between $\theta_{y,i,\tau_i} = Pr(x_i \leq \tau_i|y)$ for $y = 0, 1$. Pittman et al (2002) develop a coherent model in which the full distributions of expression levels, conditional on y , are modelled as nonparametric Dirichlet processes. Node split assessment then results in a simple Bayes’ factor to test the difference of two Bernoulli probabilities θ_{0,i,τ_i} and θ_{1,i,τ_i} at the node. This critically respects both the case-control sampling and the need for a consistent test across a range of possible thresholds τ_i , the latter ensured through the Dirichlet process models.

Bayes’ factor tests tend to be conservative, growing smaller trees than significance tests (Berger 1993). At each node, the tests rank predictor:threshold pairs and then split if the best are highly

significant. Multiple clones of the tree are generated based on multiple, significant predictor:threshold pairs, leading to “forests of trees” (Breiman 2001) which can be later weighted by relative likelihood values. Predictions averaged across trees with such weights will tend to improve predictions by respecting, and properly accounting for, tree model uncertainty.

Inference and prediction from a tree involves “branch” probabilities $\theta_{y,\tau,i}$. With conjugate Beta priors defined by the Dirichlet model, posteriors for these are also Beta. Moving down a tree through a series of branches, the product of a series of branch probabilities defines the likelihood ratio for $y = 0/1$ in the terminal leaf, and this can be converted to a predictive probability for future cases in that leaf on specifying a prior $Pr(y = 0)$. Posterior simulation easily generates estimates and posterior intervals for these predictive probabilities, and these can be mixed or resampled across trees.

Examples in breast cancer phenotyping (ER prediction, recurrence prediction) and cardiovascular disease-state prediction, appear in the above references (and at www.cagp.duke.edu). We stress the utility of aggregate, metagene predictors for their value in both dimension reduction and signal extraction – a common “signature” in a group of genes that may plausibly relate to an underlying biological pathway should be better estimated this way, while multiplicities are reduced. The particular construction of metagenes is simple, and will likely improve as new methods for identifying underlying common factors in high-dimensional data evolve (West 2003). The predictive value of the metagene tree approach is assessed using both internal cross-validation and, in one study, prediction of completely new data, with excellent results. The analyses also define sets of genes for further study – genes defining and associated with the metagenes that dominate the building of predictive models. Full discussion appears in the above references and supporting material.

4. Survival Tree Models

Suppose now y is a survival (death, cancer recurrence, etc.) time outcome, possibly right-censored, and we aim to model the survival distribution $p(y|x)$ for multiple gene, metagene and clinical variables in the high-dimensional vector x . Typical survival studies involve prospective sampling and so tree construction is more traditional in considering splits at nodes based on differences in outcome distributions $p(y|x_i)$ within any node. Our analysis (Pittman et al 2003) builds Weibull models. For a specified, global Weibull shape parameter, we can transform the data to exponential, analyse the data and build trees with exponential survival distributions, and then transform back to the original scale for predictions of new cases. For any given set of trees we are able to compute relative likelihood values, which define a joint likelihood function over trees and the Weibull parameter jointly. Then repeating this analysis across a grid of values of the shape parameter allows us to focus in on highly weighted trees and Weibull shape parameter combinations. This understood, we can now focus on tree model building with exponential distributions.

We grow and fit trees using a Bayesian approach that is analogous to the forward selection of trees in the binary case above. At any node, a candidate predictor:threshold pair $\{x_i, \tau_i\}$ organises the data in that node into two subgroups according to whether $x_i \leq \tau_i$ or $x_i > \tau_i$, and the assessment compares evidence for or against a difference in survival distributions between the two subgroups. Under the exponential model, we need a prior on the two exponential means, and a prior on the common mean if the distributions do not differ. We utilize conjugate gamma priors, and nest the null hypothesis (one common exponential) within the alternative (two different exponentials) by assuming the same gamma prior in each case. The resulting, easily computed Bayes’ factor is then the association measure for the specific predictor:threshold combination. An upper threshold on this Bayes’ factor defines the level for splitting a node based on any predictor:threshold combination. Also, multiple predictor:threshold pairs may define highly significant splits so leading to cloned copies of trees that are then split on different variables. This again generates many trees, and the resulting forests are evaluated via a final computation of approximate relative likelihood functions.

The gamma priors within each node use a fixed prior mean but treat the shape parameter as uncertain and node specific. In a node, the shape parameter is estimated via empirical Bayes' using the partition of the sample into two exponentials and the resulting marginal MLE for the shape parameter. This has two key aspects: first, it permits "borrowing strength" across the two subgroups to estimate this key parameter; second, it allows for differing prior gamma shape parameters at different nodes in each tree, thus flexibility in responding to varying degrees of uncertainty as we move down the tree. Generally, higher nodes will be easier to split based on greater numbers and diversity of data, and so will tend to warrant lower gamma prior shapes.

In one tree, inference at a leaf involves the posterior predictive distribution of future survival times for cases whose predictor variables lead to that leaf. These predictive distributions are Pareto (gamma mixtures of exponentials) that are then transformed via the Weibull shape parameter. To represent predictions across many candidate trees, we use simulation: sample tree models according to the posterior probabilities, i.e., (simply, currently) normalised relative likelihoods, then sample the implied unique Pareto distribution for a candidate future sample, based on the predictor profile of that case, in the chosen tree. Repeating this leads to a Monte Carlo sample from the predictive distribution that represents both within tree (Pareto) uncertainties and, potentially critically, uncertainty across tree models. These samples can be summarised to produce point and interval estimates of survival probabilities at any chosen set of time points, and profiles of the full predicted survival distributions.

5. An Example of Survival Tree Models: Breast Cancer Recurrence

An extensive study in application to breast cancer recurrence is described in Pittman et al (2003). This includes a number of explorations of specific trees that draw from both metagene and clinical data predictors. Overall predictive accuracy, evaluated via cross-validation, is very good and dominates traditional approaches. The x data include clinical information (lymph node counts, ER status, tumour size, etc) and values on 498 metagenes. The latter were constructed in the full set of 159 samples. From quantile normalised gene expression from Affymetrix U95a arrays, we apply k-means correlation-based clustering (following an initial screen to remove genes varying at low levels), and define the metagenes as the dominant principal component within each cluster. The idea is to extract multiple patterns as candidate predictors while reducing dimension and smoothing out gene-specific noise. The focus is not on clustering to deliver biologically interpretable groupings, though genes related to cluster patterns that have predictive value are of interest.

Here, 97 patients with "high" values on a key metagene, Mg440, identified in Pittman et al (2003) as associated with recurrence, are the focus. Survival means recurrence-free following surgery. We generated multiple trees using both metagenes and clinical predictors (including lymph node status, ER status, and treatment status). One example, the high likelihood tree in Figure 2, is typical; each node contains the predictor:threshold which created the node split, the number of sample individuals in the node, and the posterior predictive probability of 4-year recurrence free survival for that subpopulation. Figure 3 shows a snapshot of predictions for cases with more than 4 years of follow-up. These are honest predictions: each case was held out, the entire tree model generation re-performed based on the remaining 96 cases, and the hold-out case predicted. From the full predicted survival functions, this looks at just 4 year values. Predictive accuracy is high, consistent with our larger study. Several clinical variables, including treatments, appear in top trees. The approach is flexible and well suited to adapting to multiple and new, emerging forms of potentially relevant data. Finally, a number of potentially very relevant biological connections are made with some of the key metagenes arising in likely trees, including genes associated with growth factor signaling and immunological responsiveness (Pittman et al 2003), suggestive of plausible directions for biological investigation. Current studies aim to generate such follow up, as well as to substantially expand sample numbers to refine the tree model analysis.

