Bayesian CART Model Search: Rejoinder
Author(s): Hugh A. Chipman, Edward I. George, Robert E. McCulloch
Source: *Journal of the American Statistical Association*, Vol. 93, No. 443 (Sep., 1998), pp. 957-960
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2669836
Accessed: 22/01/2010 10:32

Denison, Mallick, and Smith 1998b for a comprehensive description of a Bayesian MARS model.)

ADDITIONAL REFERENCES

Denison, D. G. T. (1997), "Simulation-Based Bayesian Nonparametric Regression Methods," unpublished Ph.D. thesis, Imperial College, London.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998a), "A Bayesian CART Algorithm," *Biometrika*, 85, 363–377.

——— (1998b), "Bayesian MARS," unpublished manuscript submitted to *Statistics and Computing*.

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics*, 19, 1–141.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Rejoinder

Hugh A. CHIPMAN, Edward I. GEORGE, and Robert E. MCCULLOCH

We are very grateful to the discussants for providing additional insights and commentary on our procedure. Many of their comments clarify aspects of our work and draw attention to important issues. We are especially pleased that one of the main themes in the discussions concerns potential improvements to Bayesian CART. Indeed, our intention in writing this article was to spur further development of the Bayesian approach to CART by illustrating its potential. Clearly, much more remains to be done. We take the opportunity here to comment on some points raised, clarify some issues, and speak of a broader context. Our responses are organized mainly in the order of issues raised by the discussants.

4. DENISON, SMITH, AND MALLICK

4.1 One Long Run Versus Multiple Restarts

Despite the fact that our Markov chain Monte Carlo algorithm and the one proposed by Denison, Smith, and Mallick (DSM) tend to get stuck in local neighborhoods of posterior modes, DSM advocate the use of a constrained burn-in period followed by one long run. Although they concede that "restarts will inevitably lead us to find better structures," they favor the single-run approach because it speeds up the analysis, appears to find good models, and yields approximate posterior inference "conditional on the form of the top nodes." Indeed, it is impressive that the CART model they obtained for the breast cancer data is so similar to the one that we found with multiple restarts. (Nothing remotely like this tree was found by greedy algorithms.)

Although we appreciate the potential of the DSM strategy, it is not foolproof and should be used with caution. First of all, holding the chain up during the burn-in period will not necessarily prevent the search from going "down poor blind alleys," because the best trees will not necessarily have the most probable top node structure. Furthermore, output from a single run (such as DSM's Fig. 2) is not indicative of good performance by itself. As each of the multiple runs in our Figure 6 shows, this is exactly what a single bad run stuck near a local mode might look like. Finally, conditional posterior inference from a chain stuck near an inferior local mode will be misleadingly optimistic.

The use of multiple restarts helps avoid these limitations. Because the probability of getting stuck near an inferior local mode will always exist, the use of multiple restarts facilitates search over a larger and more diverse set of trees, thereby increasing the chances of finding a variety of good models. For example, Figure 1 here shows a high-likelihood 10-node tree for the breast cancer data that we also found using multiple restarts. This tree is structurally very different from the tree in DSM's Figure 1, and it is extremely unlikely that a single run would find both of these trees. Using multiple restarts would enhance, rather than limit, the value of conditional posterior inference, because it would be conditional on a larger and more diverse set. Exact conditional probabilities can easily be obtained by applying our closed-form posterior expressions to every visited models. Finally, if we wish to improve prediction by model averaging, a set of good but disparate trees originating from many restarts of a chain will provide a more stable mixture for prediction and generalization.

Although we are skeptical of the single-run strategy, we very much like DSM's idea of using a constrained burn-in period to better explore the top part of the tree. In a multiple restart scenario, this approach could be used to generate alternative starting trees before releasing the Markov chain to explore the full tree space. Because performance probably will depend on the region of constraint, it probably would be useful to vary this region in any implementation. Other potential improvements to the Bayesian CART algorithm are discussed in our response to Knight, Kustra, and Tibshirani that follows.

4.2 Prior Specifications and Simplicity

An appealing and impressive aspect of DSM's algorithm is the elegant simplicity of their prior specification, which contains only one prior parameter to be set. This allows for the nearly automatic application of their algorithm, rendering it accessible to a wide variety of statisticians. Although we have recommended some semiautomatic prior parameter choices for similar purposes, we have instead sought to emphasize the flexibility that is available using our priors. By using different prior choices, the user can guide the search to explore different regions of the model space. (For exam-

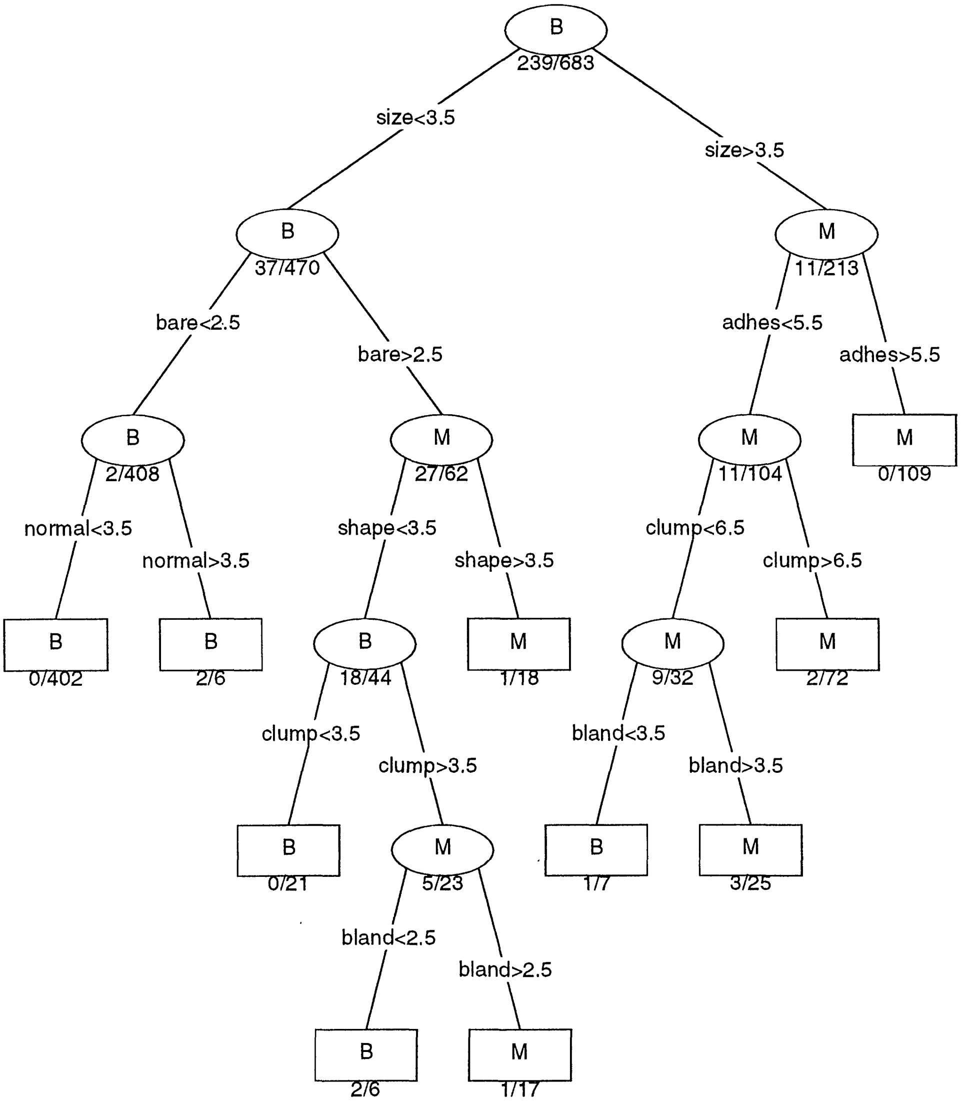

Figure 2. A Tree Found by a Different Restart of the Chain, Breast Cancer Data Example.

ple, in Chipman, George, and McCulloch 1998 we propose a hierarchical prior that effectively guides the search toward models that are appropriate for shrinkage estimation.) Finally, we are somewhat concerned about DSM's implicit use of improper priors for the terminal node parameters when different dimensional models are being compared.

4.3 Cross-Validation

We completely agree with DSM and Zhang that the in-sample misclassification rates in the breast cancer data are overly optimistic, and that out-of-sample assessments would provide a better measure of performance. We plan to report elsewhere on the use of cross-validation adjustments to compensate for potential overfitting.

4.4 Extensions

The use of Bayesian methods in complex models (such as DSM's interesting approach to MARS) is a promising growth area. Both stochastic search and posterior mixing offer advantages over traditional approaches. However, Bayesians should be aware of other procedures that offer similar advantages. Bootstrapping complex models (Breiman 1996; Tibshirani and Knight 1996) is trivial to implement and offers both improved search and mixing. Other interesting ideas include "stacking" (Wolpert 1992) and "boosting" (Freund and Schapire 1996).

5. KNIGHT, KUSTRA, AND TIBSHIRANI

5.1 Limitations of Bayesian CART

Knight, Kustra and Tibshirani (KKT) begin by raising some concerns about the limitations of our approach. First, they wonder about the robustness of our regression tree method given the assumption of normality for the terminal node distributions. Although a full answer to this question is beyond the scope of our current investigation, we are hopeful about robustness because the marginal likelihoods in (11) and (14), which are the basis of our selection criteria, are t likelihoods that are well known for their robustness properties. Note that we also accommodate considerable heteroscedasticity by allowing both the mean and variance to vary across the terminal nodes.

KKT are also concerned about the degradation to our procedure if many irrelevant predictors are added. So are we. In a sense, the algorithm is facing a variable selection problem at each node. If noise is added in the form of additional irrelevant predictors, then the chances of pursuing such noise will only increase. This is, of course, a problem for all approaches.

The extent to which irrelevant predictors are eliminated in our current implementation is determined by the rejection step of the MH algorithm. In this step, rejection probabilities based on the prior and the data are used to choose predictors. Thus priors that perform variable selection (Sec. 3.2) could help guide the search toward active variables. Another possible improvement might be KKT's suggestion to use posterior probabilities to guide the proposal distribution. This could, for example, be implemented in the GROW step by first calculating the posterior distribution of all possible splits on all possible variables at a given node (as in a greedy search) and then proposing a candidate based on a draw from this distribution. Efficient updating rules could be used to lessen the computational costs of such a step. Although such a step seems interesting, the resulting output may no longer converge to the posterior. Nonetheless, this innovation is at least worth further investigation.

5.2 Efficiency and Comparison With Bumping

The simulation results presented by KKT clearly demonstrate that our limited comparison with bumping (Sec. 6) was unduly pessimistic. Our comparison was based on a single simulation of the data, and unfortunately this specific dataset was one that made it very difficult to identify the correct model. Clearly, bumping will usually be much more effective on this problem than our article suggests.

However, KKT's reported comparisons between bumping and Bayesian CART are somewhat misleading, because one iteration of the Bayesian CART procedure is computationally much less expensive than growing an entire CART tree from a bootstrap sample. As KKT correctly surmise, the computation of the MH transition probability (19) (under normal regression tree models) requires $O(n)$ operations, which is then roughly the cost of each Bayesian CART iteration. (We say roughly, because different proposal have different costs. For example, a CHANGE or SWAP move has to check whether it makes any terminal nodes logically empty.) Given that growing a CART tree from a single resampling requires $O(nmp)$ operations, their reported median of 116 required resamplings to find the true tree is comparable to our reported median of 4,000 required iterations. This is perhaps not so surprising, given that bumping can be seen as an approximation to posterior sampling. Similarly, the 25,000 that iterations we used to improve on bumped trees correspond to far fewer than 25,000 resampled trees.

5.3 New Proposal Steps

By using an experiment to study tree variability, KKT develop two promising new proposal steps to add to the MH algorithm: the BOTTOM-UP step and the ROTATE step, both of which can be considered as extensions of our SWAP step. We included the BOTTOM-UP step in our SWAP step but only when two child nodes have exactly the same rules. KKT's BOTTOM-UP step strengthens this by allowing swapping with approximately equal splitting rules. The ROTATE step can also be thought of as a multiple SWAP step. As KKT's Figure 4 demonstrates, ROTATE can be especially useful in exposing superfluous splits to pruning. In addition to these two new steps, the way in which KKT discovered them is also very important. Having struggled ourselves to come up with new proposal steps, the idea of using systematic strategies, such as studying tree variability and clustering around local posterior spikes, may well turn out to be a key insight for the development of MCMC proposals for this and other problems.

KKT's innovations led us consider yet another extension of the SWAP step, which we call SWAP-AT-JOIN. This rule is motivated by noticing that if the correct splitting rule appears in the wrong node, then the same or similar rule is likely to appear in a different subtree as well. For example, if in the simulated data of Section 6, the first split incorrectly uses X_1, then it is likely that both children would later involve splits on the correct first rule, namely $X_2 \in \{A, B\}$ versus $\{C, D\}$. This error would be corrected by the following:

- SWAP-AT-JOIN: Identify two nodes with the same (or nearly identical) rule. Trace the lineage of these two nodes back until a common parent is arrived at. Swap the rule of the common parent with the two original children.

6. ZHANG

6.1 The Swap Step and CART Construction

As Zhang points out, the SWAP step can be very effective when used in a nonstochastic construction strategy. As of Zhang's Figures 1–3 illustrate, SWAP is most effective when used in conjunction with other tree-modifying steps, because more than simple parent–child swaps were performed to move from Figure 1 to Figure 3. Indeed, it would be interesting to consider general nonstochastic strategies that used all of the proposal steps that we have considered, including those suggested by KKT. Such strategies could be automatic or could involve structured user intervention.

Unfortunately, the complexity of CART construction can easily overwhelm such strategies. In this regard, stochastic search can be a useful alternative for exploring the CART space and identifying a diverse set of models, especially if it is guided by a meaningful posterior distribution as we have proposed. To this end, we feel that the combination of the SWAP step with stochastic search is particularly powerful. The effectiveness of the SWAP step depends strongly on the makeup of the tree under consideration. For example, SWAP will be ineffective on a CART tree that uses the wrong variables. But, when the ingredients of a good tree occur, but not in the right topology, swapping will be a useful tool.

6.2 Scientific Plausibility

Zhang emphasizes the importance of maintaining scientific plausibility over maximizing likelihood or posterior probability for choosing a model. We agree completely, and would argue that this is a strong reason for preferring an approach such as ours, which provides a diverse set of candidate models from which interpretable models can be selected. This is a further argument for preferring multiple restarts to a single long run, as we discussed in our response to DSM.

A related issue that we have considered is the development of methods for making sense of the CART output from automatic procedures (Bayesian or otherwise). For example, if hundreds of well-fitting trees were produced, then it would be useful to cluster them into meaningful groups of similar trees. We plan to report further on this work elsewhere. (See also Shannon and Banks 1997.)

6.3 Reproducibility

Although reproducibility is a comforting feature when available, users of Markov Chain Monte Carlo methods will simply have to forfeit this luxury. Useful reassurance is still provided by favorable comparisons to baselines like the greedy tree. We believe that the value of model is determined by how well it explains a phenomenon, not how it was discovered.

6.4 Exact Posterior Probabilities

To assign an exact probability to a given tree, it would be necessary to evaluate the norming constant for (17), which would in turn require evaluating the right side of (17) for *every* possible tree. Although this will usually be impractical, (17) does allow us to compute the relative probability of any two trees, which is all that is really needed for tree selection from the set of visited trees. In fact, in large problems the actual tree posterior probabilities are apt to be very small, because of the enormous size of the tree space.

ADDITIONAL REFERENCES

Freund, Y., and Schapire, R. E. (1996), "Experiments With a New Boosting Algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning*, ed. L. Siatta, San Francisco: Morgan Kaufmann, pp. 148–156.

Shannon, W., and Banks, D. (1997), "An MLE Strategy for Combining CART Models," in *Proceedings of the 1997 Symposium on the Interface: Computing Science and Statistics*, unpublished manuscript.

Wolpert, D. (1992), "Stacked Generalization," *Neural Networks*, 5, 241–259.