posterior probabilities have been settled for some time." In sharp contrast, we have strongly cautioned against such a strategy, because even with our additional SWAP step, the algorithm quickly gets trapped in a local posterior mode, after which it only moves locally (see Sec. 6). Our recommended strategy of continual restarts is deliberately designed to avoid this problem.

[Received March 1996. Revised November 1997.]

REFERENCES

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.

Buntine, W. (1992), "Learning Classification Trees," *Statistics and Computing*, 2, 63–73.

Clark, L., and Pregibon, D. (1992), "Tree-Based Models" in *Statistical Models in S*, eds. J. Chambers and T. Hastie, Belmont, CA: Wadsworth.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Hierarchical Priors for Bayesian CART Shrinkage," Working Paper 98-03, University of Waterloo, Dept. of Statistics and Actuarial Science.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (in press), "A Bayesian CART Algorithm," *Biometrika*, 85.

George, E. I. (1998), "Bayesian Model Selection," in *Encyclopedia of Statistical Sciences Update*, Vol. 3, New York: Wiley.

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Jordan, M. I., and Jacobs, R. A. (1994), "Mixtures of Experts and the EM Algorithm," *Neural Computation*, 6, 181–214.

Lutsko, J. F., and Kuijpers, B. (1994), "Simulated Annealing in the Construction of Near-Optimal Decision Trees," in *Selecting Models From Data: AI and Statistics IV*, eds. P. Cheeseman and R. W. Oldford, New York: Springer-Verlag, pp. 453–462.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–676.

Oliver, J. J., and Hand, D. J. (1995), "On Pruning and Averaging Decision Trees," in *Proceedings of the International Machine Learning Conference*, pp. 430–437.

Paass, G., and Kindermann, J. (1997), "Describing the Uncertainty of Bayesian Predictions By Using Ensembles of Models and Its Application," in *1997 Real World Computing Symposium*, Real World Computing Partnership, Tsukuba Research Center, Tsukuba, Japan, pp. 118–125.

Quinlan, J. R., and Rivest, R. L. (1989), "Inferring Decision Trees Using the Minimum Description Length Principle," *Information and Computation*, 80, 227–248.

Sutton, C. (1991), "Improving Classification Trees With Simulated Annealing," in *Proceedings of the 23rd Symposium on the Interface*, ed. E. Keramidas, Interface Foundation of North America.

Tibshirani, R., and Knight, K. (1995), "Model Search and Inference by Bootstrap 'Bumping'," technical report, University of Toronto.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1762.

Wallace, C. C., and Patrick, J. D. (1993), "Coding Decision Trees," *Machine Learning*, 11, 7–22.

Wolberg, W. H., and Mangasarian, O. L. (1990), "Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology," *Proceedings of the National Academy of Sciences*, 87, 9193–9196.

Comment

Heping ZHANG

Bayesian model selection in CART is a fruitful idea; however, what I liked most in this article is the idea of swapping, which I explain shortly. The authors emphasized the potential superiority of the Bayesian framework and the stochastic search in finding CART models. I raise the following questions:

- Is it evident that the Bayesian CART model search as put forward by the authors is helpful in finding the "best" CART models?
- What are important issues that may be resolved with Bayesian ideas?
- What are the caveats of the Bayesian CART model search?

It appears to me that the Bayesian CART model search can be useful, but the evidence to date is not convincing. To illustrate this point, I reanalyze the dataset used by the authors in their Section 7. Before doing that, I clarify the criteria on which I will judge the quality of the trees. In the authors' Figure 1, the misclassification cost is calculated using the resubstitution method. Breiman et al. (1984) discussed in depth that this cost estimate is overly optimistic. In other words, we should not get too excited by the low resubstitution cost estimate. Despite this caution, to continue from the authors' discussion, let me also base our discussion on the resubstitution cost estimate. Obviously, the more trees one examines, the better chance of ending up with an improved tree. It is not too hard to see that the Bayesian model selection searches over a far greater number of candidate trees than the greedy algorithm. Thus Bayesian model selection should presumably do a better job than the greedy algorithm.

Figure 1 displays a tree built from the program that I designed using the entropy criteria as a splitting principle (see Zhang and Bracken 1996 and Zhang, Holford, and Bracken 1996). The program is available on the Web at http://peace.med.yale.edu:8000/pub/rtree, or by anony mous ftp. This program can build a tree automatically (like the construction of Fig. 1) or interactively with the user (like the construction of Figs. 2 and 3). The tree in Figure 1 has seven terminal nodes and results in 23 counts of misclassification if node 13 is classified as "benign" and 22 counts of misclassification if node 13 is classified as "malig-

Heping Zhang is Associate Professor of Biostatistics, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520-8034 (E-mail: heping.zhang@yale.edu).

Journal of the American Statistical Association
September 1998, Vol. 93, No. 443, Theory and Methods

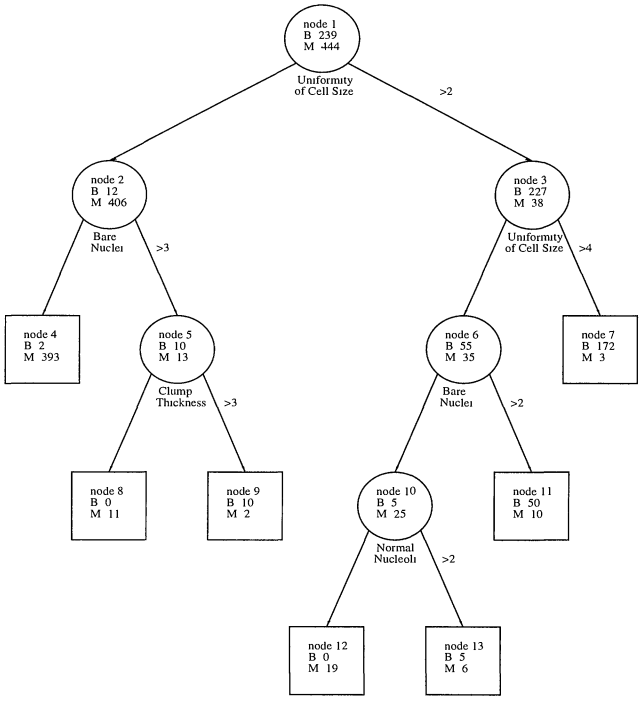

Figure 1. A Tree Obtained Automatically. The numbers of benign (B) and malignant (M) cases are given inside each node.

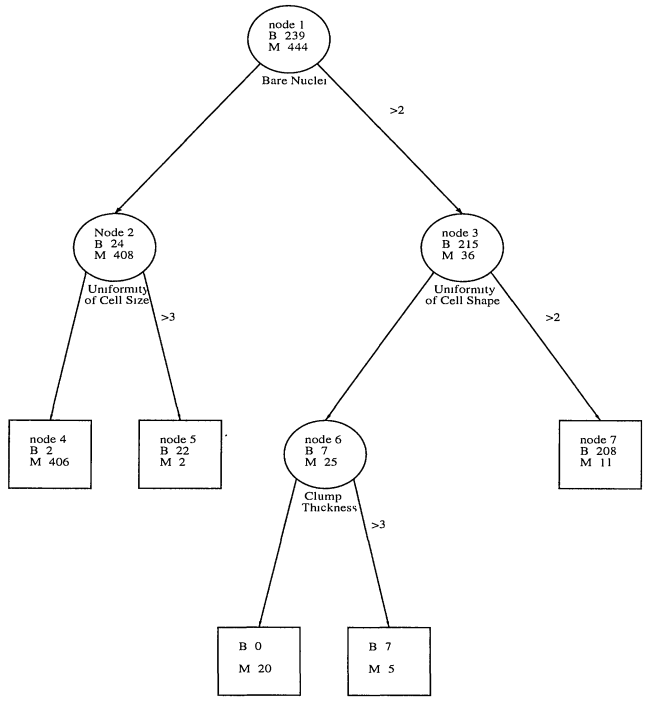

Figure 2. A Tree After One Swap. The numbers of benign (B) and malignant (M) cases are given inside each node.

nant." It is interesting to realize that nodes 1 and 3 are split by the same variable (uniformity of cell size), and nodes 2 and 6 are divided by another variable (bare nuclei). Using the swapping idea stated in the article, it seems natural to swap the splitting variables between nodes 1 and 2. Figure 2 results from this swap, where node 1 is partitioned by the number of bare nuclei and node 2 by uniformity of cell size. The tree has the same basic profile as that in the authors' Figure 1, and it makes use of the same set of splitting variables. The differences are in the cut-off values for the splits and the locations that call for the splitting variables. This new tree misclassifies 20 cases. If we use the swap idea again and split node 3 in Figure 2 by clump thickness, we end up with the tree in Figure 3. The only difference between my tree in Figure 3 and the tree in the authors' Figure 1 is the cut-off value of clump thickness in node 3. Both trees collect the same number of misclassifications of 18.

What is the point of this exercise? It seems to support the idea that swapping is really worthwhile. It may not be generally true, but in this particular case it led to a certain degree of improvement in terms of the resubstitution misclassification cost every time it was applied. It is important to note that I did not use any stochastic search, as the authors did. The implication from this application is that one may "improve" the quality of trees by swapping splitting variables without the intensive stochastic search. Thus it would be useful if the authors could clarify what really helps: swapping, stochastic searching, or both?

It is statistically desirable to maximize the likelihood or posterior probability. When a statistical method is applied in the real world, the most important matter is scientific plausibility. The greedy algorithm of Breiman et al. (1984) and Bayesian model selection presented in the article build trees automatically. This is a very nice feature in some situations, but not always. In my own experience and in my communications with other users, the first automated tree does not necessarily make scientific sense. Alternative, competitive trees must also be presented to allow scientists to select a tree that has reasonable scientific justification.

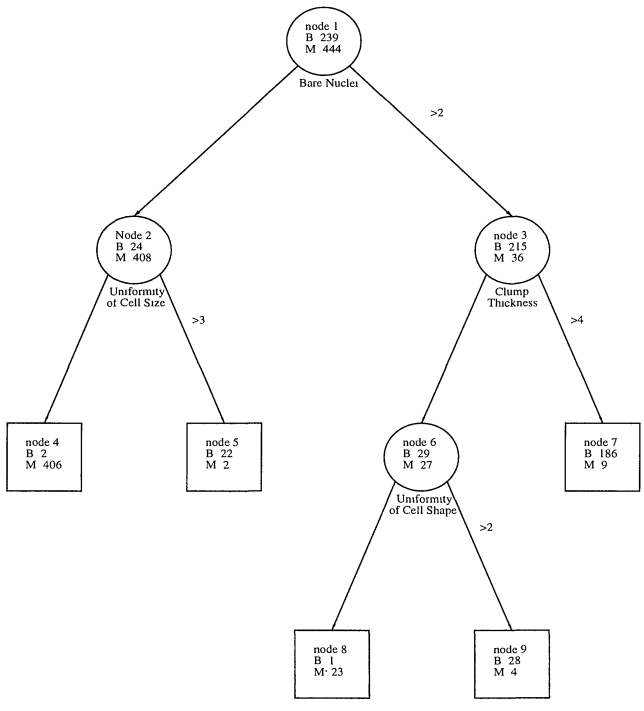

Figure 3. A Tree After Two Swaps. The numbers of benign (B) and malignant (M) cases are given inside each node.

The Bayesian model selection seems to be a potentially useful approach in generating several trees that are competitive in a purely statistical sense, one of which may distinguish itself from the rest by its scientific plausibility.

Tree stability is another important and difficult issue that the Bayesian model selection may help. It would be useful if the authors could provide some insight into this issue. The authors point out that it is nearly impossible to determine exactly which trees have the largest posterior probability. But is it possible to assign the posterior probability to a given tree? A tree with a high likelihood or posterior probability will give the user a sense of security when adopting the tree.

From a user's viewpoint, if I take the dataset such as the one in Section 7 and use the authors' program, how likely is it that I will find the same tree? Using the greedy algorithm, the size of a tree is subject to random variation in pruning; however, the basic tree structure is determined in the tree-growing step. Thus to a certain extent, the tree is reproducible. Users could be concerned, frustrated, and possibly even embarrassed when unable to reproduce their trees or trees published by others.

Although the misclassification costs of the tree in Figure 1 of the article and the one in my Figure 3 are slightly lower than those of the trees in my Figures 1 and 2, it does not mean that the improvement would be sustained if a test sample were applied. In addition, because Section 7 deals with a real application, it would be more interesting to address scientific issues rather than a small amount of reduction in the misclassification cost. For example, does the tree in the authors' Figure 1 make better clinical sense than that in their Figure 10?

ADDITIONAL REFERENCES

Zhang, H. P., and Bracken, M. B. (1996), "A Tree-Based Two-Stage Risk Factor Analysis of Spontaneous Abortion," *American Journal of Epidemiology*, 144, 989–996.

Zhang, H. P., Holford, T. R., and Bracken, M. B. (1996), "A Tree-Based Method in Prospective Studies," *Statistics in Medicine*, 15, 37–50.

Comment

Keith KNIGHT, Rafal KUSTRA, and Robert TIBSHIRANI

In this interesting article, the authors propose a Bayesian method for searching through the space of regression or classification trees. As the authors make clear, although their proposal is based on the Markov chain Monte Carlo method, it is not able to generate trees from a full posterior distribution of trees. Rather, it suggests trees of high probability and hence is useful for finding "good" CART trees. The idea is to improve on the greedy tree search used by the CART algorithm, which can produce poor trees that represent suboptimal local minima. In this sense their proposal is Bayesian technique for assisting a non-Bayesian method of analysis.

One concern we have about the authors' method involves the many parametric assumptions made at both the data and the prior stages. The CART method is nonparametric in flavor—specifically, the squared error criterion used for regression trees works reasonably well for non-Gaussian data. In contrast, the authors make Gaussian assumptions in equations (3) and (4)—how well will the method perform for nonnormal data? Another concern is the simple rule p_{RULE} for choosing which predictor to split. The rule chooses uniformly among the available predictors. If we add many predictors unrelated to the response to our dataset, will this rule not degrade? It would seem to spend most of its time visiting trees that split on useless predictors. It seems that one needs a rule that chooses predictors according to their posterior probability given the data, but this may be difficult to compute.

In the remainder of this discussion, we revisit the comparison with "bumping" and make some proposals for improving the authors' procedure.

1. COMPARISON WITH BUMPING

As the authors note, Tibshirani and Knight (1996) proposed a method called "bumping" for finding better local minima. It works by applying the modelling procedure (such as the CART tree-growing algorithm) to a set of B bootstrap samples—that is, samples of the same size as the original dataset, drawn with replacement from it. This produces a collection of B models—in the present example, B CART trees. We then choose the tree that fits the original dataset the best, using as our criterion squared (training) error or any other convenient measure. We always include the original dataset as one of the bootstrap samples, so that the original model can compete with the others. Of course is not fair to compare the training error of models of different complexity. Hence we must first decide on the model complexity (in the present case, tree size), either subjectively or by an adaptive method such as cross-validation.

Keith Knight is Associate Professor, Department of Statistics, Rafal Kustra is a graduate student, Department of Public Health Services, and Robert Tibshirani is Professor, Department of Public Health Sciences and Department of Statistics, University of Toronto, Ontario, Canada. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Journal of the American Statistical Association
September 1998, Vol. 93, No. 443, Theory and Methods