Statistical Science
2006, Vol. 21, No. 1, 30–34
DOI 10.1214/088342306000000079

Rejoinder: Classifier Technology and the Illusion of Progress

David J. Hand

I would like to thank the discussants for some very stimulating comments. Being only human, I am naturally pleased when others produce evidence or arguments in support of my contentions, but being a scientist, I am also pleased when others produce evidence or arguments against my proposals (although I may have to take a deep breath first), since this represents the scientific process in action.

I should first make one thing clear: I agree with Professor Friedman that substantial advances have been made in recent years. Indeed, in my paper I remarked that "developments such as the bootstrap and other resampling approaches ... have led to significant advances in classification and other statistical models." However, what I question is whether the advances, when taken in the context of real practical problems, are as great as is often claimed—the recognition of the limitations of the new methods to which Professor Friedman refers.

Professor Friedman agrees with my three points that the improvements of newer methods over older ones are less than those of the older ones over still older ones, that the evidence favoring the superiority of new methods is often suspect and that the new methods fail to tackle important problems. I draw the conclusion from these points that progress is not as great as is imagined. Professor Friedman draws the conclusion that low lying fruit is easier to gather, that initial validation of new methods should be more rigorous and that much work remains to be done. Perhaps, then, we are really broadly in agreement—only perhaps I am describing a half empty glass (the new classification tools are not as wonderful as they are claimed), while Professor Friedman is describing a half full glass (some classification tools represent advances over the older ones).

I admit that I did criticize error rate as a performance measure and then used it in the examples. Since most performance comparisons of classifiers use error rate, this seemed justifiable, and I believe that my conclusions will generalize to other performance measures. For example, I agree that in some two-class problems it is the rank order of the estimated class 1 membership probabilities which matters and that modern methods may well be able to estimate this more accurately than older methods. However, surely my points about population drift, class definition uncertainty and so on still apply and, of course, my point that people often use one criterion to fit a model and another to evaluate it applies even more strongly.

In fact, this point about people using different criteria manifests itself at a higher level when Professor Friedman and I examine my Table 1. I see the proportion of reduction of error rate achieved by the best method which can be achieved by discriminant analysis, whereas Professor Friedman sees the ratio of the error rates. I see a large initial improvement so that subsequent improvements are relatively small; he sees a large reduction in the proportion remaining. Back to the half full/half empty glasses again. We are both right, of course, although perhaps the different perspectives are valuable for different uses. For example, I agree with Professor Friedman's example of the zip code classifier—and here the ratio of error rates might be a sensible measure—but (I would imagine) this is a problem in which the distributions are fairly static. In other problems, the distributions will change rapidly and I can imagine many contexts when I would not want to place too much trust in a reduction of error rate by a factor even as large as 10, if it corresponded to a change from a starting point as small as 0.001 to an even smaller one of 0.0001. A slight shift in the shapes of the distributions might induce sufficiently large changes in error rate so as to make this change irrelevant.

My regression example in Section 2.1 was merely intended as an additional illustration of the fact that the sequential nature of modeling means that typically later improvements are smaller than early ones. I am suggesting that the first, relatively crude, models will generally yield greater marginal improvements in predictive power than the later models. This is the low hanging fruit phenomenon—athough, as noted below and as Professor Stine illustrates, there are exceptions.

I am glad Professor Friedman agrees so strongly with Section 5 of the paper, on the difficulties of obtaining