Statistical Science
2006, Vol. 21, No. 1, 27–29
DOI 10.1214/088342306000000042

Comment: Classifier Technology and the Illusion of Progress

Robert A. Stine

It is my pleasure to contribute to the discussion of this paper. David Hand has the credibility one needs to write such an article and not have it dismissed out of Hand. Along with publishing numerous papers and books on classification and data mining, he "works in the trenches" with real data. His contributions to credit modeling are particularly well known and respected, and his knowledge of that domain reaches far deeper into the substance than the casual illustration often chosen to show off a new methodology. He is a fascinating lecturer and I have learned a great deal by listening carefully to his ideas. When he writes that claims of the superiority of neural networks and support vector machines "fail to take account of important aspects of real problems," I have to stop and think about my own research and experiences.

The thrust of Hand's paper is the argument that most recent developments in classification, say anything since Fisher's linear discriminant function, offer little benefit in practice. The mismatch between theory and practice dwarfs incremental claims for superiority established in theorems. For instance, theory that shows that a support vector machine classifies better than a simple linear model is an "illusion," bordering on sophistry.

I have a great deal of sympathy for this point of view, but I doubt that many statisticians will change what they do after reading this paper. I agree with many of his criticisms, but I am already in the choir. I suspect that it will take quite a bit more to convince others, particularly along the lines of proposals for what ought to be done. Consider the impact of Tukey's "The future of data analysis" (Tukey, 1962). After chastising the field for its preoccupation with "optimization in terms of a precise, similarly inadequate criterion," Tukey proposed alternatives, including exploratory data analysis and robust methods. Forty years later, Hand's criticisms echo his concerns.

Robert A. Stine is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: stine@wharton.upenn.edu).

Hand presents a range of criticisms of modern classifiers. I find it useful to organize my discussion by grouping them into two clusters:

- Creeping incrementalism
- Square pegs in round holes.

Let me start with the first of these.

Creeping incrementalism. Hand argues that concerns for optimality emphasize tiny improvements that are dwarfed by other issues in real applications. He argues that the first predictor or the most simple of models finds most of the structure. Adding bells and whistles contributes little more than complex window dressing, and the advantages are illusions that disappear during the application. The argument is analogous to saying that linear Taylor series make pretty good approximations to most functions; generally, you do not need those messy, higher order terms. I certainly agree that simple models—or at least simple methodologies—take you a long way. Dean Foster and I wrote a paper to make just this point when mining financial data: with a few adjustments, stepwise linear regression can predict bankruptcy as well as elaborate trees (Foster and Stine, 2004).

A convincing argument for preferring simpler models requires careful discussions of applications. Given the depth of his experience, I had expected Hand to offer a rich portfolio of examples that demonstrate the failures of complex models. Instead, he relies more on an idealized example (one of equally correlated predictors) and a summary of fitted models to selected data sets from the repository at UC Irvine. One has to be careful basing arguments on made-up examples, because it is too easy to turn the examples around. With equally correlated predictors, the first one or two predictors capture most of the signal, with diminishing benefits left to the others. Although I have had similar experiences modeling real data, it is all too easy to make up normal models in which later variables appear to explain the most variation. For example, define

$$\begin{aligned} X_1 &= \tau Y + \varepsilon_1 + \varepsilon_2, \\ X_2 &= \tau Y + \varepsilon_1 - \varepsilon_2, \\ X_3 &= \tau Y - \varepsilon_1 + \varepsilon_3, \\ X_4 &= \tau Y - \varepsilon_1 - \varepsilon_3, \end{aligned} \quad \text{where } Y, \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1). \tag{1}$$