Comment

Elie Bienenstock and Stuart Geman

According to the authors, this paper has three principal goals: "informs a statistical readership about Artificial Neural Networks (ANNs), points out some of the links with statistical methodology and encourages cross-disciplinary research...." It seems to us that the authors have been spectacularly successful with regards to the first two of these goals, and it is likely that this paper will do much to further stimulate the already active scientific exchange between the statistics and neural modeling communities.

As Cheng and Titterington made clear, neural networks, at least the very popular examples reviewed in their paper, are not really new inasmuch as they represent variations on common statistical themes, especially nonparametric and semiparametric estimation and classification. Furthermore, Cheng and Titterington suggest that the tie to real neurons may be somewhat tenuous (we will amplify on this shortly). Nevertheless, despite this dubious biological connection and strong ties to already well-studied statistical methods, this field has attracted wide attention from within the government (principally the Department of Defense but also other branches including the Department of Commerce) as well as many sectors of industry. It has drawn many top science students at our top schools. In the meantime, many statistics departments complain that it is hard to find first-rate graduate students.

We would like to use this discussion to speculate about the reasons behind the fantastic growth of the neural modeling field, especially in light of the close ties to well-studied areas of statistics which have themselves been received with substantially less enthusiasm. There are many reasons for the remarkable popularity and visibility of neural networks. We will propose a few and suggest that some of them may be based partly on misconceptions.

THE APPEAL OF BRAIN MODELING

The endeavor is nearly irresistible: building models and machines possessing a measure of human intelligence, working through the puzzles of perception and cognition and "explaining" the brain. Indeed, many researchers in the neural modeling community believe that the kinds of networks discussed by Cheng and Titterington are meaningfully connected with biology, providing a starting point from which we can begin to organize and understand the overwhelmingly complex anatomical and physiological data, and from which new kinds of theoretically-directed biological experiments will emerge. Still, most neural modelers would agree that these attempts are nothing more than the crudest of approximations not to be taken seriously as models of real neurons or real neuronal interactions at the level of any important detail. Cheng and Titterington have already remarked that "it is clear that the brain does not learn by the generalized delta rule." It is also clear that there is very little in the way of feedforward networks in the brain (virtually all substantial pathways are reciprocated making it clear that the dynamics is not that of a feedforward network) and that the real equations of synaptic modification are a good deal more complicated than a Hebbian or gradient-descent rule. In short, ANNs are hardly neural.

THE APPEAL OF "GENERALIZATION"

Model-free generalization has served as a kind of Holy Grail in neural modeling: begin with a more-or-less *tabula rasa* (blank slate, or, in statistical parlance, "nonparametric") architecture and a realistically-sized training set for some challenging classification or estimation task and devise a learning rule powerful enough to discover the regularities and invariants that would extrapolate good performance beyond the training data. Such a device might be used to "beat the stock market" or solve the automatic target recognition (ATR) problem which has resisted many years of expensive R&D effort. But statisticians know that generalization (good performance on samples not in the training set) depends almost entirely on the extent to which the training set is representative, and/or the structure of the problem happens to accommodate the models used. It is too much to expect statistical methods to "discover," by themselves, complex and nontrivial structure such as the structure

Elie Bienenstock is Visiting Associate Professor (on leave from CNRS, Paris, France) and Stuart Geman is Professor, Division of Applied Mathematics, Brown University, Box F, Providence, Rhode Island 02912.

that defines classes of objects, invariant to lighting, shading, texturing, rigid and nonrigid shape deformations and viewing perspectives. The situation with pre-segmented hand-written numerals is quite special: this is a small class of essentially one-dimensional structures for which very large and comprehensive training sets are available.

Of course, the problem of recognizing handwritten numerals is an important one, and there are many other problems of equal importance which are equally amenable to neural network and related statistical approaches. However, it has been observed many times that for such problems simple nearest-neighbor methods (or variations on that theme) typically perform nearly as well (and often better) than neural networks [see, for example, the thorough experiments by Ripley (1993)]. Evidently, in these cases "generalization" is mostly a matter of *interpolation.*

We have argued elsewhere (Geman, Bienenstock and Doursat, 1992) that for many of the more ambitious problems for which neural networks have been proposed (such as ATR, unconstrained handwriting recognition or learning complex motor maps for robot arms with multiple degrees of freedom), the choice of a suitable statistical method may ultimately play only a minor role. The more substantial challenge may prove to be the choice of appropriate *representations*, in particular, representations in which generalization can, in fact, be viewed as a matter of interpolating a sufficiently rich but reasonably-sized training set. We would argue, for example, that unconstrained object recognition will require the development of representations that are already nearly invariant to pose, shape, lighting, etc., and that "learning" such representations from examples is nearly impossible with realistic training sets.

Cheng and Titterington remark that two principal steps in treating a practical problem are (i) the specification of an appropriate architecture, and (ii) network training from examples. We would like to suggest that substantial progress on the more ambitious problems for which neural networks have been proposed will require a shift in emphasis from issues of training to issues of architecture—which is to say, modeling.

PROBLEM SELECTION

Cheng and Titterington began their paper with a list of currently used—in some cases about-to-be-used—applications of ANNs. The list is impressive, and one could no doubt add more items to it, such as the various applications to high-energy physics (e.g., see Denby, 1993) to mention but one area. The fact that ANNs have been successfully applied to work with real data for substantial problems in speech synthesis (NETtalk), speech recognition, character recognition and robotics has certainly contributed much to their appeal. However, it should be mentioned that there is the tendency to somewhat exaggerate the successes. After about ten years of intense activity in the field, the number of concrete industrial applications is still rather limited. Many "applications" are really *demonstrations*, and it is often the case that neural nets are outperformed by (less general) *ad hoc* solutions. This, for example, is the situation with NETtalk, as Cheng and Titterington have pointed out.

PACKAGING

The importance of an appealing presentation cannot be ignored, even in science. Cheng and Titterington rightly remark that ANNs are sometimes perceived, from the perspective of statisticians, as "familiar entities" with a representation that is "usually pictorial." Although the last two words appear in parentheses in the paper, they could actually be taken as one of the main take-home messages. What is a radial-basis-function ANN if not a kernel method for regression *with a picture*? Figure 8 is the picture of a two-layer perceptron, but this is nothing more than a particular nonlinear regression model. In fact, wording itself can play a substantial role. Contrast the very intuitive notions used in the definition of Boltzmann machines—hidden units; clamped and unclamped dynamics; Hebbian synaptic plasticity—to the rather unappealing statistical terminology (to quote again from the paper): "a version of the iterative proportional fitting procedure used in analyzing multiway contingency tables." For that matter, also consider the phrase "Boltzmann machine" against "semiparametric estimation via maximum likelihood."

We would like to conclude by observing that, despite these reservations, there is little doubt that the popularity of ANNs has had, and continues to have, a very positive effect on scientific research. It has brought together scientists from diverse disciplines to work on important and interesting problems (numerous prominent theoretical physicists, mathematicians, computer scientists and biologists have adopted the field as a kind of second career), and it has done much to advertise the enormous potential of statistics for addressing a host of modern, "high-technology," problems. Cheng and Titterington's paper should be welcomed as further encouragement to this kind of important cross-disciplinary research.

ACKNOWLEDGMENTS

This work was supported by Army Research Office contract DAAL03-92-G-0115 to the Center for Intelligent Control Systems, NSF Grant DMS-92-17655 and Office of Naval Research Contract N00014-91-J-1021.

Comment

Leo Breiman

Cheng and Titterington have most commendably brought developments in the neural network field to the attention of statisticians. It is a notable public service. Since their title is worded "...A Review from a Statistical Perspective", room is left for other statistical perspectives.

When I first heard about neural networks some years ago, I was put off by what I considered to be the hype about doing things the way the brain does. The going propaganda seemed to be that here was a set of procedures modeled after the brain that did a miraculously accurate job in a wide variety of tasks. The functioning of these procedures was coded in esoteric language based on terms borrowed from brain mechanisms. The whole thing was reminiscent of the artificial intelligence publicity a decade or two ago.

But in going to neural network meetings, reading and refereeing their articles and talking to many practitioners over the last five years, my opinion has changed. The neural network community consists of different segments. Some are concerned with constructing mathematical network models of the brain. Others are concerned with networks as mathematical entities, that is, their connectedness, dynamics, etc. Probably the largest segment consists of the people doing work on pattern recognition and other predictive problems.

1. THE CHARACTERISTICS OF THIS LATTER COMMUNITY

They are *not* a neural network community. They use any methodology that works on their problems. Often, they use CART or MARS. They experiment with nearest neighbor methods, separating surfaces gotten by using linear programming, radial basis functions, hidden Markov chains, etc. New methodologies are constantly proposed, and many of these have little resemblance to standard neural networks. Unfortunately, much of the original, and now anachronistic, terminology is retained giving misleading impressions about what is going on.

They are very pragmatic and problem oriented. In fact, the field is better defined by the nature of the problems they work on then by any particular methodology. Typical problems are speech recognition and handwritten character recognition. The range of problems is characterized by high dimensional complex data, often with very large sample sizes (10^4 to 10^7). The goal is to find accurate predictors in classification, regression and time series.

Often, the methodology they use is hand-tailored to the problem they are working on. In this respect, the neural network technology is attractive in that the network and the number of internal nodes can be tinkered with and optimized for the problem. But other methods are employed if they give better results.

Their bottom line is the error rate on the relevant data set. Proposed new methodologies are judged in terms of their error rates on banks of known data sets. But there is little systematic research into the circumstances under which some methods work better than others. This may be because the work is so oriented toward particular problem solving and tailored methodologies.

The people involved are, by background, computer scientists, engineers and physical scientists. They are generally young, energetic and highly computer literate. They have the further good fortune not to have any formal statistical training so that they feel no compulsion to engage in the futile games of modeling data or in endless asymptotics. What they have borrowed from statistics is very slight.

There are important cultural differences between the statistical and neural network communities. If a statistician analyzes data, the first question he gets asked is "what's your data model?" The NN practitioner will be asked "what's your accuracy?" In

Leo Breiman is Professor, Department of Statistics, 367 Evans Hall, University of California, Berkeley, California 94720.