White Paper

It is a Mistake to…. Rely on One Technique

by John Elder

Founder & CEO

October 2013

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 3 (of 11) of a series by the author on the Top 10 Data Mining Mistakes, drawn from the *Handbook of Statistical Analysis and Data Mining Applications*.

"To a little boy with a hammer, all the world's a nail." All of us have had colleagues (at least) for whom the best solution for a problem happens to be the type of analysis in which they are most skilled! For many reasons, most researchers and practitioners focus too narrowly on one type of modeling technique. But, for best results, one needs a whole toolkit. At the very least, be sure to compare any new and promising method against a stodgy conventional one, such as linear regression (LR) or linear discriminant analysis (LDA). In a study of articles in a Neural Network journal over a 3-year period (about a decade ago), only 17% of the articles avoided this mistake and the one in last issue (Focusing too much on Training). That is, five of every six refereed articles either looked only at training data or didn't compare results against a baseline method, or made both of those mistakes. One can only assume that conference papers and unpublished experiments, subject to less scrutiny, are even less rigorous.

Using only one modeling method leads one to credit (or blame) it for the results. Most often, it is more accurate to blame the data. It is unusual for the particular modeling technique to make more difference than the expertise of the practitioner or the inherent difficulty of the data — and when the method will matter strongly is hard to predict. It is best to employ a handful of good tools. Once the data is made useful — which usually eats most of your time – running another algorithm with which you are familiar and analyzing its results adds only 5-10% more effort. (But, to your client, boss, or research reviewers, it looks like twice the work!)

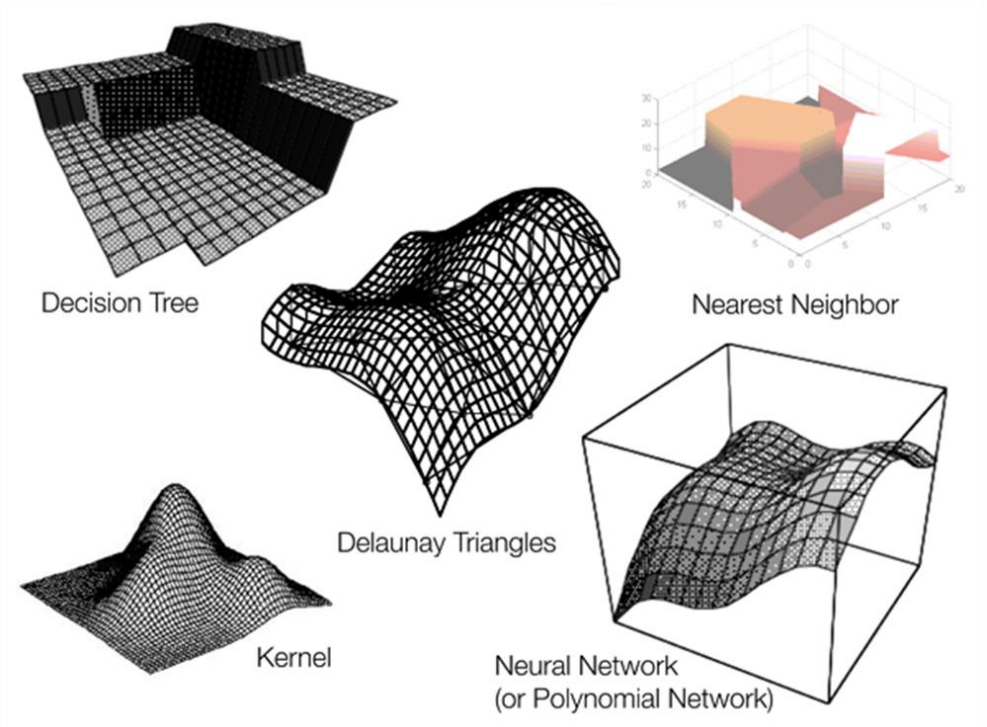

Figure 1

The true variety of modeling algorithms is much less than the apparent variety, as many methods devolve to variations on a handful of elemental forms. But, there are real differences in how that handful build surfaces to "connect the dots" of the training data, as illustrated in Figure 1 for five different methods – *Decision Tree, Polynomial Network, Delaunay Triangles* (which I invented), *Adaptive Kernels*, and *Nearest Neighbors* — on (different) two-dimensional input data. Surely some surfaces have characteristics more appropriate than others for a given problem.

Figure 2 (from work done with Stephen Lee) reveals this performance issue graphically. The relative error of five different methods – *Neural Network, Logistic regression, Linear Vector Quantization, Projection Pursuit Regression,* and *Decision Tree* – is plotted for six different problems from the Machine Learning Repository.[1] Note that "every dog has its day"; that is, that every method wins or nearly wins on at least one problem.[2] On this set of experiments, *Neural Networks* came out best, but how does one predict beforehand which technique will work best for your problem? [3] Best to try several, and even use a combination (as in Chapter 13 of the *Handbook*, and to be covered in more detail in installment 11 of 11).

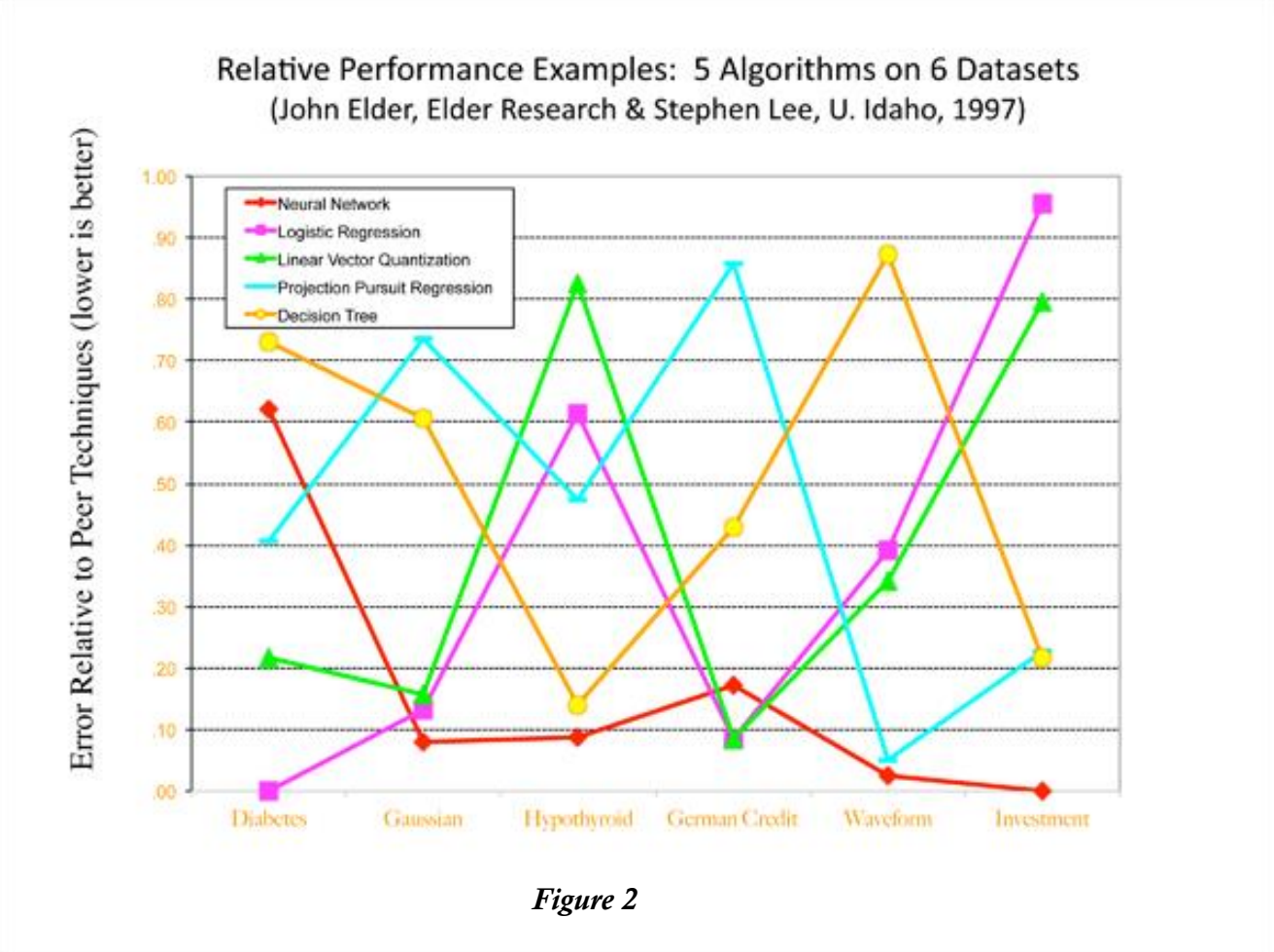

Figure 2

About the Author

Dr. John Elder, Founder and CEO of Elder Research, leads the largest and most experienced data science consulting firm in the U.S. For 20 years, the team has applied advanced analytics to achieve high ROI for investment, commercial and security clients in fields from text mining and stock selection, to credit scoring and fraud detection. John has Engineering degrees from Rice and the University of Virginia, where he's an adjunct professor. He's authored innovative tools, is a popular keynote speaker, and has chaired International Analytics conferences. Dr. Elder served 5 years on a panel appointed by President Bush to guide technology for National Security. He has co-authored three books (on data mining, ensemble modeling, and text mining), two of which won Prose "book of the year" awards.

---

[1] The worst out-of-sample error for each problem is shown as a value near 1 and the best as near 0. The problems, along the X-axis, are arranged left-to-right by increasing proportion of error variance. So, the methods differed least on the Pima Indians Diabetes data and most on the (toy) Investment data. The models were built by advocates of the techniques (using *S*implementations), reducing the "tender loving care" factor of performance differences. Still, the UCI ML repository data are over-studied; there are likely fewer cases in those datasets than there are papers employing them!

[2] When I used this colloquialism in a presentation in Santiago, Chile the excellent translator employed a quite different Spanish phrase, roughly "Tell the pig Christmas is coming!" I had meant every method has a situation in which it celebrates; the translation conveyed the concept on the flip-side: "You think you're something, eh pig? Well, soon you'll be dinner!"

[3] An excellent comparative study examining nearly two dozen methods (though nine are variations on *Decision Trees*) against as many problems is (Michie, Spiegelhalter, and Taylor, 1994, reviewed by me for JASA 1996). Armed with the matrix of results, the authors even built a decision tree to predict which method would work best on a problem with given data characteristics.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673