The Interface'98 Conference – a Resource for KDD

John F. Elder IV, Ph.D.
Elder Research
1006 Wildmere Place
Charlottesville, VA 22901
804-973-7673
www.dataminingag.com

Elder@datamininglab.com

ABSTRACT

The annual *Symposium on the Interface between Computer Science and Statistics* is the statistical conference likely to be of greatest interest to quantitative practitioners of KDD**.**

Keywords

Data Mining, Statistics

1. INTRODUCTION

Begun 30 years ago, the Interface symposium was inspired by early (albeit over-promising) research in Artificial Intelligence, and the possibility of new modes of inference based heavily on computation. In the decades since, it has been a leading forum for the introduction of ground-breaking inference techniques and algorithms. Though, like most conferences, it tends to be dominated by academics, and though most presenters are statisticians, it has done well in attracting significant participation both from other disciplines and from industry.

This note describes the distinctives of that conference and summarizes some of the interesting strands of research presented at last May's meeting in Minneapolis, "*Dimension Reduction, Computational Complexity and Information*". The next meeting, with the focus "*Models, Predictions, and Computing*", is June 9-12 in Schaumburg, Illinois (near Chicago) http://www.math.niu.edu/Interface99/ It features inexpensive, 1-day short courses on *Data Mining* (by myself), *Repeated Measures* (Ed Vonesh), and *Graphical Methods for Categorical Data Analysis* (Heike Hofmann and Antony Unwin). The keynote speaker is Leo Brieman, and banquet speaker is Stephan Wolfram. Plans for 2000 have the conference in New Orleans, and there are hopes to co-locate it with KDD in 2001.

2. INTERFACE DISTINCTIVES

For those who value practical results, a chief virtue of the Interface is the participation of many practical and algorithmic statisticians – i.e., "constructivists". Several pioneers who frequently contribute to the Interface have made the effort to speak the language of other disciplines and might (or should) be familiar to Machine Learning and KDD researchers. These include Jerry Friedman, Leo Brieman, Trevor Hastie, Rob Tibshirani, Daryl Pregibon, Brian Ripley, David Hand, David Scott, Grace Wahba, Ed Wegman, Andreas Buja, etc. – collectively responsible for an "alphabet soup" of innovative methods.

From a contributor's standpoint, two properties of the Interface are notable: 1) one's paper is due a month <u>after</u> the conference (allowing more recent research to be described), and 2) virtually all proposed talks are accepted. One might guess the latter property diminishes the quality of the (4-5 parallel) sessions, but actually, the level is high. I believe this is due to the strong self-filtering of the statistical community. In other words, the caution and exactness generally characterizing statisticians – which traits have often been thought to impede creativity and boldness from the perspective of other communities -- nevertheless serves to insure that when something new is proposed, it is likely to be trustworthy.

3. HIGHLIGHTS FROM 1998

In 1998, the algorithms discussed included Trees, MARS, splines, LASSO, SIR, wavelets, local polynomials, kernels, simulated annealing, grand tour, projection pursuit, and ridges. Commercial applications dealt with included credit scoring, internet traffic, health care, the environment, Y2K, DNA, remote sensing, and motion control. And, among the topic areas covered were re-sampling (bootstrapping), visualization, graphical interfaces, web-based statistical education tools, dimensionality reduction, complexity measurement, model combination (bundling), and extreme values. A companion article by Arnold Goodman in this newsletter gives more of an overview of the last *Interface*. What follows is a brief summary of a couple of the research results – representative of *Interface* topics – most interesting to this author.

3.1 Measuring Model Complexity

Perhaps the primary issue with automated model induction (the heart of Data Mining) is deciding when to stop. To avoid overfit, it is standard to constrain complexity, under the Occam's-razor assumption that simplicity is a virtue in conflict with training accuracy. But, to control something, one must be able to measure it. Perhaps the problems with this strongly-held assumption – recently enthusiastically critiqued by Pedro Dominqos in his award-winning talk at KDD last year [2] – have much to do with improperly assessing true complexity.

Typically, one counts the parameters being fit, weights this linearly, and compares against fitting improvements (using AIC, MDL, or other information-theoretic criteria). This is the right approach for Linear Regression, but not for nonlinear approaches, whose terms typically are more powerful (e.g., a knot location in MARS empirically has the strength of about 3 linear terms) and sometimes much less influential (e.g., weights in an under-trained neural network). (Moreover, as the last example makes clear, effective strength can change with training effort, given no change in the model structure.)

Jianming Ye presented an intriguing alternative metric, Generalized Degrees of Freedom, which finds the sensitivity of the fitted values to perturbations in the outputs [5]. This allows complexity to be measured and compared in an entirely experimental manner, suitable for very complex, multi-stage

Data Mining processes. The complexity depends not only on the hypothesized model, but on the algorithm, data, and difficulty of the problem, and is measured in a way requiring few assumptions. As an example, a decision tree with 19 nodes, built using 100 noisy observations with 10 variables, was found to use the equivalent of 79 degrees of freedom -- not the 19 that might normally be assigned. Should this metric prove to work, one could fairly compare competing models with diverse pedigrees, and better guard against overfit.

3.2 Model Variability and its Uses

A key statistical insight – not fully appreciated by researchers in most fields – is that the data one has is just a part of all that is possible. Hence, the saying that "Being a statistician means never having to say you're sure"; some certainty is always reserved in one's judgements out of deep respect for (fear of?) the variability which permeates reality.

Data Mining algorithms are so flexible (to address different types of systems) that often a tiny change in the input set, observation count, or guiding parameters, etc. will lead to a seemingly great change in the model selected. This is disturbing to those trying to interpret a model – a property sought by many, and a reason for the popularity of methods like Decision Trees and Rule Induction. But is a change in *form* (model structure) important if it leads to only a small change in *function* (estimation surface)? This debate, and others related to the measurement and meaning of model variability, was energetically engaged at the Interface, with interesting experiments sure to follow.

Lastly, as oft-discovered in recent years, multiple models have uses beyond the estimation of variability; they can act as a committee of experts to achieve improved performance. This hot research topic of "bundling" or combining competing models (.e.g, bagging, boosting, arcing, stacking, etc.) – which some statisticians have been pondering for years in a Bayesian framework – received close attention as a promising technique. KDD researchers will surely welcome the rigorous statistical examination (e.g., [4]).of these heuristic ideas – an examination they are helping to provoke by simply showing that the ideas *work*, and hence are worthy of sustained attention.

4. AN INVITATION

Over the decades, Statistics has been *the* major source of useful data-based techniques (e.g., [3]), though the discipline has performed poorly at extending and exploiting those results. Jerry Friedman says "If a statistician gets a good idea, he writes a paper. If a computer scientist does, he starts a company." As an engineer, I can get away with characterizing my Engineering and CS colleagues as "not letting ignorance about a subject keep us from addressing it". Daryl Pregibon, head of AT&T's venerable statistical group, puts it more charitably: "Data Miners are fearless". This boldness and creativity is, on balance, a virtue; still, the more cautious approach of statisticians has its benefits. The groups certainly have the potential to make strong allies.

At Interface '98, computer scientist Thomas Deitterich gave a masterful summary of the major currents in Machine Learning [1], which was very well-received by the attendees.[1] In '99 Usama Fayyad will describe leading KDD issues. Such reviews, or interesting current results from related fields, are very welcome at the *Interface between Computing Science and Statistics*, and I urge researchers to contribute. Also, I highly recommend KDD practitioners attend. Let us expand the dialog capable of leading to reliable progress in our related fields. The extra effort should make us even more capable of our shared goal of finding useful patterns in data.

REFERENCES

(The *1998 Interface Proceedings* have sadly not yet arrived; so, where necessary, I've noted recent published versions of the work described.)

[1] Deitterich, T. (Winter, 1997). Machine Learning Research: Four Current Directions. *AI Magazine*.

[2] Domingos, P. (1998). Occam's Two Razors: The Sharp and the Blunt. Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, New York, AAAI Press.

[3] Elder, J. F. IV & D. Pregibon (1996). A Statistical Perspective on Knowledge Discovery in Databases, Chapter 4 in *Advances in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI/MIT Press.

[4] Friedman, J. H. (February 1999). Greedy Function Approximation: A Gradient Boosting Machine. Stanford University, Dept. of Statistics, Technical Report.

[5] Ye, J. (March 1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association* **93**, no. 441, pp. 120-131.

About the author:

John Elder is Chief Scientist of a data mining consulting firm in Charlottesville, Virginia (www.dataminininglab.com). He has 15 years experience developing and applying adaptive, data-driven techniques to practical problems - at an engineering consulting firm, an investment management company, Rice University, and the University of Virginia. Dr. Elder has written and spoken widely on pattern discovery and has authored a handful of influential data mining programs. He is associate editor of *Statistics and Computing*, is on the board of the Interface Foundation, and chairs the Adaptive and Learning Systems Group of the IEEE-SMC Society.

[1] The five topics highlighted were ensembles of classifiers, coupling classifiers, scaling up ML algorithms, reinforcement learning, and learning with stochastic methods.