Knowledge Discovery and the Interface of Computing and Statistics

Arnold Goodman
The UCI Center for Statistical Consulting
University of California, Irvine
Irvine, CA 92697-5105
agoodman@uci.edu

John Elder IV
Elder Research
1006 Wildmere Place
Charlottesvi-lle, VA 22901
elder@dataminingLab.com

Abstract

Brief but stimulating overviews are presented to place knowledge discovery and data mining (KDD) in the data analysis and research cycles:

- o An introduction with an historical perspective
- o Research success: critical factor checklist, to improve likelihood of research success
- o Bridging the KDD-statistics gap, with key relevant questions and issues as a first step
- o The Interface '98 papers relevant for KDD

Introduction and Historical Perspective

KDD is that proactive new area of information technology driven by enormous data bases with significant knowledge buried deep inside them. It's objective is to make useful sense out of data.

The data is not experimental but opportunistic. It just happens to be there with alluring potential.

It can be diverse, heterogeneous, overwhelming and maybe even non stationary in space and time. Necessity demands data be analyzed all together.

Analysis of the data is application oriented and driven by computation. Probability and statistics are there to help but not to hinder getting results.

Although this situation is outside the bounds of statistical tradition, it is not outside the boundary of statistical techniques and statistical thinking. It may be observed that KDD is reminiscent of the real beginnings of statistics: there was data to be mined before there was a theory to guide it.

KDD did not really intend to become statistics. However, it is not only using statistics, but also contributing to statistics. It is on the interface of computing and statistics, and there is much to be achieved by both data miners and statisticians, in bridging the technique and thinking gap between these two essentially different yet similar fields.

It was a 1965 UCI seminar by Arthur Samuel, on teaching the computer to play checkers, that inspired Arnold Goodman to conceive the whole interface of computer science and statistics. It is composed of employing computers for statistical problems, using statistics in computer problems, and utilizing both of them on those significant problems of other important knowledge areas.

Interface '67 had sessions on computational linguistics, artificial intelligence, applications within the Interface and computer simulation.

In 1972, Barry Merrill at State Farm Insurance became the first commercial customer of SAS, when he mined the very first data warehouse of SMF records from IBM mainframe computers.

Decision trees were provided to KDD by Leo Breiman, Jerry Friedman, Richard Olshen and Charles Stone (four statisticians) in their 1984 book on the extremely useful CART technique.

Both Interface '97 and Interface '98 Keynote Addresses dealt with KDD: Jerry Friedman said that statistics is no longer the only data game in town, and David Rocke said that the algorithm is the estimator. KDD poses great opportunities for statistics, as well as great challenges to statistics.

KDD-97 appeared at Interface '98, Interface'98 is appearing at KDD- 98, and KDD-98 will itself appear at Interface '99. A relationship is being developed between KDD and Interface, to lead toward increasing their interaction and perhaps a collocated KDD-01 and Interface '01 meeting. Interaction will produce progress, if not synergy.

The insightful comments by Kay Anderson and Tom Dietterich on the essential nature of KDD, and by Padhraic Smyth on some significant differences between KDD and statistics are gratefully acknowledged here.

Research Success: Critical Factor Checklist

Extend inferences to plans, evaluations & values.

How Checklist Works. The summary comments:

.00 Weight factors according to their importance
.00 Add weights of those factors accomplished
.00 Give a partial weight for a portion done
.00 Score might simulate a "probability"

Where to Start. Weights for factors:

.01 Constraints from data collection
.01 Preprocessing or transformations
.01 Prioritized information objectives
.01 What is to be included & what not
.01 Where to learn & where to validate

How to Start. Weights for factors:

.02 Data questions
.02 Problem needs
.02 Prior knowledge
.02 Problem structure
.02 Meaningful models
.02 Applicable methods
.02 Useful visualizations
.02 Experimental designs

When to Stop. Weights for factors:

.02 Iteration stability
.02 Result consistency
.02 Data support excess
.02 Problem requirements
.02 Learning validation tips
.02 Maybe sequential analysis

What to Infer. Weights for factors:

.03 Clusters & regions
.03 Distributions & trees
.03 Relationships & models
.03 Comparisons & contrasts
.03 Indications of any changes

What to Do Next. Weights for factors:

.03 Posing questions & issues to resolve
.03 Using different techniques on the data
.03 Using the techniques on different data
.03 Developing the meaning for inferences
.03 Presenting the inferences to customers
.03 Generating added methods & theory

What to Evaluate. Weights for factors:

.02 Assumption/data accuracy & compatibility
.02 Size variations & strength of relationships
.02 Truth & meaningfulness of all inferences
.02 Consistency & sensitivity of inferences
.02 Importance vs statistical significance
.02 Confirmation or criticism of model
.02 "Distance" of inferences from data
.02 Client's understanding of inferences
.02 Use of inferences in problem context
.02 Quality vs timeliness of the inferences
.02 Cost-effectiveness of existing inferences
.02 Cost-effectiveness of additional inferences

What the Objective Is. Weights for factors:

.01 Objective was data mining
.01 Objective was data refining
.01 Objective was data defining
.01 This influences when to stop
.01 This influences what to infer
.01 This influences what to do next
.01 This influences how to evaluate

What Resulting Value Is. Weights for factors:

.01 To expense, income & competitive position
.01 To quality, time & customer service
.01 To understanding & advances

Bridging the KDD-Statistics Gap

Data Stages. The key questions & issues follow.

- What can statisticians contribute to progress without detracting from required efficiency?
- What are the most important & pressing data mining issues statisticians can contribute to?
- What & how can data miners contribute to the data stages & to statistics in particular?
- What & how can statisticians contribute to the data stages & to data mining explicitly?
- Think about having statisticians evaluate the statistical consequences of KDD software.
- Think about developing e-mail lists of those statisticians desiring to collaborate on KDD.
- Think about developing e-mail lists of data miners desiring to collaborate on statistics.
- Think about all the data mining, data refining & data defining stages while in every stage.
- Think about the requirements for transition from current stage to the following stage(s).
- Act appropriately for the current stage, but not to detriment of the following stage(s).

Data Mining. Key questions & issues follow.

- What can statisticians contribute to efficiency without detracting from the effectiveness?
- Think about getting ready & aiming to fire before firing without getting ready or aiming.
- Think about capitalizing on the estimates & testing research before variation searching.
- Think about happening often & happening almost always before believe happening once.
- Think about targeting later correlations & relationships before targeting coincidences.
- Think about the required symbolism & its terminology before being buried in acronyms.

Data Refining. Key questions & issues follow.

- How can both data miners & statisticians cooperate to transition beyond data mining?
- Think to explore with vigor & explain with rigor before exploiting with too much zeal.
- Think about desired information & desired knowledge before finding desired meaning.
- Think about technology engine & science engine before utilizing mathematics engine.
- Think interesting/informative & insightful/ institutional prior to indicative/ingenious.

Data Defining. Key questions & issues follow.

- What can data miners contribute to being better without detracting from productivity?
- Think about reducing results & generalizing implications before understanding methods.
- Think about bringing size under control & process conclusions before truth of a theory.
- Think about seeking change & utilizing change before just assessing the change.
- Think about the needed flexibility & synthesis before letting discipline reign.
- Think about preceding pictures & formulas before relying on merely concepts & jargon.

Research Stages. Some key issues follow.

- Think about all of the research stages & their characteristics while operating in every stage.
- Think about the requirements for a transition from current stage to the following stage(s).
- Act appropriately for the current stage, but not to the detriment of the following stage(s).

Where to Start. Some key issues follow.

- Think about what statisticians might actually contribute to information objectives priority.
- Think about what statisticians might actually contribute to where learn & where validate.

How to Start. Some key issues follow.

- Think about what statisticians might actually contribute to building a problem's structure.
- Think about what statisticians might actually contribute to considering experiment designs.

When to Stop. Some key issues follow.

- Think about what statisticians might actually contribute to analysis not exceeding support.
- Think about what statisticians might actually contribute to utilizing sequential analysis.

What to Infer. Some key issues follow.

- Think about what statisticians might actually contribute to relationship & model inference.
- Think about what statisticians might actually contribute to indications of change inference.

What to Do Next. Some key issues follow.

- Think about what statisticians might actually contribute to use different techniques on data.
- Think about what statisticians might actually contribute to developing inference meaning.

What to Evaluate. Some key issues follow.

- Think about what statisticians might actually contribute to inference truth/meaningfulness.
- Think about what statisticians might actually contribute to seeing result distance from data.

What Resulting Value Is. Key issues follow.

- What techniques can statisticians introduce to clarify, expedite & guide valuation process?
- Think about what statisticians might actually contribute to find the thinking & techniques for identifying value when it is visible & for developing value when it is not easily visible.

Interface '98 Papers Relevant for KDD

How to Start. Internet-measurement papers are:

- Vern Paxson, "Statistical Challenges in Analyzing the Internet" -- pooling diverse & heterogeneous data to find islands of stability
- John Quarterman, "Visualization of Internet Data" -- using geographical maps & graphs to measure internet quality of service from data

- Walter Willinger, “Finding Order within Chaos” -- using wavelets to identify & detect scaling properties in nonstationary space/time

When to Stop. Tree-based methods papers are:

- David Banks, “Maximum Entropy Models for Graph-Valued Random Variables” -- model selection to reflect application metrics
- Hugh Chipman, Edward George & Robert McCulloch, “Making Sense of a Forest of Trees” -- metrics to see archetypes & clusters
- William Shannon, “Averaging Classification Tree Models” -- using maximum likelihood & consensus estimates of “mean & variance”

What to Infer. Software technology papers are:

- R. Douglas Martin & Michael Sannella, “An Iconic Programming Interface for S-Plus and Mathcad” -- using component technology to integrate data & computing across packages
- David Wishart, “Exploiting the Graphical User Interface in Statistical Software: The Next Generation” -- with filtering, wizards, tutoring, visuals, model design, multiple windows, what-if’s & internet linkages
- David Woodruff, “Heuristic Search Algorithms: Applications in & of Statistics” -- for combinatorial analysis in multivariate problem robustness & cluster analysis

The four distribution & five tree papers are:

- David Marchette & Carey Priebe, “Alternating Kernel and Mixture Density Estimation” -- a semi-parametric approach
- Michael Minnotte, “Higher Order Histosplines: New Directions in Bin Smoothing” -- an extension to multivariate, nonparametric & boundary effect applications
- David Scott, “On Fitting and Adapting of Density Estimates” -- using moments of bins & polynomial patches for kernel estimates
- Sung Ahn & Edward Wegman, “A Penalty Function Method for Simplifying Adaptive Mixtures Density Estimates” -- reduces the complexity for mixtures of normal densities
- Andreas Buja & Yung-Seop Lee, “Criteria for Growing Classification and Regression Trees” -- better interpretation by splitting data based on the better performing of subsamples
- Steven Ellis, Christine Waternaux, Xinhua Liu & J. John Mann, “Comparison of Classification and Regression Trees in S-Plus and CART” -- CART has tree evaluation, error estimation & subsampling while S-Plus has better computing, graphics & interpreting
- Douglas Hawkins & Bret Musser, “One Tree or a Forest? Alternative Dendrographic Models” -- comparing trees plus their generating rules via membership & metrics
- Padraic Neville, “Growing Trees for Naive Bayes and Score Card Models” -- title says it
- S. Stanley Young & Andrew Rusinko III, “Data Mining of Large High Throughput Screening Data Sets” -- extending recursive partitioning, for relating chemical structure to biological activity, for many-many variables

The six relationship & model papers are:

- Julian Faraway, “Data Splitting Strategies for Assessing Model Selection Effects on Inference” -- performance is no better than using all the data for selection & inference
- Hakbae Lee, “Exploring Binary Response Regression Based on Dimension Reduction” -- sliced inverse regression, sliced average variance estimation & covariance differences
- Wei Pan, “Bias/Variance Tradeoff in Combining Subsample Estimates for a Very Large Data Set” -- the title essentially says it
- Armin Roehrl, “Fast, Portable, Predictable and Scalable Bootstrapping” -- using bulk synchronous parallel computing to bootstrap
- Terry Therneau, “Penalized Cox Model in S-Plus” -- with smoothing splines & frailty
- Jimmy Ye, “On Measuring and Correcting the Effects of Data Mining and Model Selection” -- generalized degrees of freedom framework for comparing the “costs” of alternative modeling & mining techniques

Reference. Weisberg, S. ed. 1999. *Proceedings of Interface ‘98: 30th Symposium on the Interface of Computing and Statistics.* Forthcoming.