Data Snooping, Dredging and Fishing: The Dark Side of Data Mining A SIGKDD99 Panel Report

David Jensen
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610

jensen@cs.umass.edu

ABSTRACT

This article briefly describes a panel discussion at SIGKDD99.

Keywords

Overfitting, SIGKDD99, Panels

1. INTRODUCTION

It is unusual for a conference to publicize the "dark side" of its topic, but KDD99 did just that. A panel at the conference, entitled "Data Snooping, Dredging and Fishing: The Dark Side of Data Mining," dealt with some pitfalls of data mining and how to avoid them. The panel was organized by Halbert White, Professor of Economics at the University of California, San Diego. In addition to White, panelists included Edward Leamer, Professor of Economics at the University of California, Los Angeles, and David Jensen, Research Assistant Professor of Computer Science at the University of Massachusetts, Amherst.[1]

The panel addressed the unique statistical challenges produced by searching a space of models and evaluating model quality based on a single data sample. Such search is common in knowledge discovery. Indeed, the term "data mining" is sometimes used pejoratively to describe such work, particularly when an analyst has searched over a large model space without adjusting for such a search or testing the resulting model on new data. Failure to adequately adjust for the statistical effects of search in large model spaces can cause a variety of problems, including excessive structure in induced models, suboptimal model construction, and vast overestimates of models' accuracy.

To many audience members, the theme of the panel was not a new one. Work in statistics on "specification searches" and "multiple comparisons" has long explored the implications of data mining, and statisticians have also developed several adjustments to account for the effects of search. Work in machine learning and knowledge discovery related to "overfitting" and "oversearching" has also explored similar themes, and researchers in these fields have also developed techniques such as pruning and minimum description length encoding to adjust for the effects of search.

However, this "dark side" of data mining is still largely unknown to some practitioners, and problems such as overfitting and overestimation of accuracy still arise in knowledge discovery applications with surprising regularity. In addition, the statistical effects of search can be quite subtle, and they can trip up even experienced researchers and practitioners.

2. EXAMPLES

The panelists came from differing academic backgrounds, but they shared a common concern for the pitfalls of data mining. Each presented some of their favorite (and particularly egregious) examples of such pitfalls.

Jensen discussed a pathology of many algorithms for constructing classification trees [7]. Over a wide variety of data sets and pruning techniques, such algorithms introduce substantial amounts of unnecessary structure into trees, and the amount of unnecessary structure increases with the size of the training set. Indeed, these algorithms will produce large trees even when given completely random data sets. Other examples discussed by Jensen included algorithms for finding "hidden messages" within large bodies of text. Such algorithms have been applied to find messages supposedly encoded within religious texts, but they can also be applied (with equal success) to find messages encoded within any text — including such documents as the Microsoft license agreement and Tolstoy's *War and Peace*.

Leamer presented several examples of implicit and explicit searches that have produced apparently useful predictors of stock market performance. For example, professional football results appear to predict overall market performance rather well: 17 NFC victories correctly predicted a market rise and 5 AFC losses correctly predicted a market decline. In the period examined, only three failures were reported. Similarly, Elizabeth Taylor's marriages in 1951, 1953, 1958, 1960, 1965, 1976, and 1977 all coincided with stock market gains. Finally, the GDP-adjusted Standard & Poors Index coincides well with the number of 45-50 year olds in the country. Over the period 1946-1997, the correlation of the two variables is 0.927. Of course, in each of these cases, an analyst should consider the size of the model space searched to find the given association. In the latter case, for example, the given relationship is the maximum correlation among 16 different age ranges.

White discussed examples drawn from his own analysis of technical trading rules in financial markets [8]. Technical trading rules are often evaluated against past stock market data, but those evaluations are rarely corrected for the "survivorship bias" that results in implicit search over the entire space of possible technical trading rules. Only rules that work well receive sustained attention from the investment community; rules that do not are either rejected by an individual analyst, or are eventually rejected by the wider community of traders (Malkiel [6] notes a

[1] One of the panelists, David Jensen, is the author of this article. Contact information for all panelists is provided at the end of the article.

similar bias in performance reports of many mutual funds). White and his colleagues examined the performance of trading rules published in 1986 on data from the period 1987-1996, and found that their performance was not significantly different than a default strategy of holding cash. In another paper, White and his colleagues examine "calendar effects" that are presumed to predict systematic changes in stock price on specific days of the week, month of the year, etc. [9]. After correcting for the size of the search space (nearly 9,500 different possible calendar effects), they conclude that no calendar rule appears to be capable of outperforming the benchmark market index over the period 1897-1996.

3. PERSPECTIVES

After discussing these examples of data mining pitfalls, the panelists delved into the problems of data mining in more detail.

Leamer characterized data mining problems with a two-by-two matrix. Columns separate problems by whether context affects the analysis. Rows separate problems by whether the analysis is predetermined and programmable or whether it is sequential with substantial human input. He noted that while the majority of statistical theory has been developed for context-independent and programmable problems (a set of assumptions Leamer called "asymptopia"), many of the most interesting problems are context-dependent and sequential. His book [5] develops statistical theory for these cases, which are still only rarely examined by statisticians.

Jensen discussed several pathologies of knowledge discovery algorithms, each caused by failure to correct for the statistical effects of search. These include overfitting, oversearching, and attribute selection errors [4]. He couched his discussion in terms of "multiple comparison procedures" (MCPs). An MCP generates multiple items, scores each item based on a data sample, and then selects the item with the maximum score. MCPs are a common component of knowledge discovery algorithms. He showed briefly how MCPs affect the sampling distribution of the selected item, and how that effect leads to the pathologies above.

4. SOLUTIONS

The panelists discussed a variety of approaches to correct for the statistical effects of searching large model spaces, including:

- *New data and cross-validation:* A very common approach is to obtain new data or to divide an existing sample into two or more subsamples, using one subsample to select a small number of models and the other subsamples to obtain unbiased scores. Cross-validation is a related approach that can be used when the process for identifying a "best" model is algorithmic.
- *Sidak, Bonferroni, and other adjustments:* Several relatively simple mathematical adjustments can be made to statistical significance tests to correct for the effects of multiple comparisons. These have been explored in detail within the statistical literature on experimental design (e.g., [2]). Unfortunately, the assumptions of these adjustments are often restrictive.
- *Resampling and randomization techniques:* Many of the most successful approaches are based on computationally-intensive techniques such as randomization and resampling. For example, White's bootstrap approach [11,8] is based on resampling. Similarly, randomization tests have been employed in several knowledge discovery algorithms [3,1].

5. CONCLUSIONS

The panelists sometimes differed in background, perspective on the problem, and suggested solutions, but all concurred that serious statistical problems are introduced by searching large model spaces, and that unwary analysts and researchers can still fall prey to these pitfalls.

6. CONTACT INFORMATION

Halbert White
http://weber.ucsd.edu/~mbacci/white/
http://www.quantmetrics.com/
mailto:hwhite@weber.ucsd.edu

Edward Leamer
http://www.anderson.ucla.edu/acad_unit/bus_econ/bios/leamerbio.htm
mailto:edward.leamer@anderson.ucla.edu

David Jensen
http://www.cs.umass.edu/~jensen/
mailto:jensen@cs.umass.edu

7. REFERENCES

[1] Frank, E. and Ian H. Witten. Using a permutation test for attribute selection in decision trees. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann. 152-160. 1998.

[2] Hochberg, Y. and A. Tamhane. *Multiple Comparison Procedures*. Wiley. 1987.

[3] Jensen, D. Induction with Randomization Testing: Decision-Oriented Analysis of Large Data Sets. Doctoral dissertation. St. Louis, MO: Washington University. 1992. http://eksl-www.cs.umass.edu/ ~jensen/papers/dissertation.ps

[4] Jensen, D. and P.R. Cohen. Multiple comparisons in induction algorithms. *Machine Learning* 38(3). http://eksl-www.cs.umass.edu/ ~jensen/papers/mlj99/jensen-cohen-ml.ps

[5] Leamer, E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley. 1978

[6] Malkiel, B. *A Random Walk Down Wall Street*. Norton. 1996.

[7] Oates, T. and D. Jensen. Large datasets lead to overly complex models: An explanation and a solution. In *Proceedings of The Fourth International Conference on Knowledge Discovery and Data Mining*. 1998. 294-298. http://www-eksl.cs.umass.edu/ papers/oates-kdd98-Datasets.ps

[8] Sullivan, R., A. Timmermann, and H. White. Data-snooping, technical trading rule performance, and the bootstrap. UCSD Department of Economics Discussion Paper #97-31. December 1997. ftp://weber.ucsd.edu/ pub/econlib/dpapers/ucsd9731.ps.gz

[9] Sullivan, R., A. Timmermann, and H. White. Dangers of data-driven inference: The case of calendar effects in stock returns. UCSD Department of Economics Discussion Paper

#98-16. June 1998. http://weber.ucsd.edu/Depts/Econ/Wpapers/Files/ucsd9816.pdf

[10] Westfall, P. and S. Young. *Resampling-Based Multiple Testing*. Wiley. 1993

[11] White, H. A Reality Check for Data Snooping. QRDA, LLC Technical Report #98-01. May 1997.

About the author:

David Jensen is Research Assistant Professor of Computer Science at the University of Massachusetts, Amherst. His current research is on relational knowledge discovery, the statistical properties of knowledge discovery algorithms, and learning for multiagent coordination. Professor Jensen served on the program committees of KDD97, KDD98, and KDD99. He received his doctoral degree in Engineering from Washington University in St. Louis in 1992. From 1991 to 1995, he was an analyst with the Office of Technology Assessment, an analytical agency of the United States Congress.