SIAM Minisymposium
on
Challenges for Data Miners, Statisticians and Clients

May 2, 2003
Held in conjunction with
the
Third SIAM Data Mining Conference, San Francisco

Organizers
Arnold Goodman, Center for Statistical Consulting, UCI
Padhraic Smyth, Computer and Information Science, UCI

Abstract
It is productive to view the life-cycle stages for data as:

- Birth in collection of data, perhaps huge or complex
- Infancy in databases of managed storage and access
- Childhood in interesting patterns and their findings
- Youth in modeled predictions and their conclusions
- Maturity in knowledge and action recommendations

There is value hidden within the data, and a real attempt should be made to maximize the amount captured during each stage of data life. The value captured at each stage depends upon how much is captured in preceding stages.

Fundamental Challenge is for data miners, statisticians and clients involved in serious data mining to recognize and accept their critical dependence, and for each of them to widen his focus until collaboration is advanced.

Productivity Challenge is for the mined findings to be transformed into provocative conclusions and then into critical action or decision recommendations, which work almost all (not only some) of the time and account for uncertainties outside (as well as inside) the database.

Evaluation Challenge is to balance the efforts spent on analysis inside the process with efforts spent on serious evaluation outside the process in client's environment, difficult though it might be to actually accomplish this.

In addition, the Symposium will explore such key issues as:

- What unutilized or under-utilized statistical techniques might be productive in data mining, why should they be, and why haven't they not been utilized enough
- Applications where statisticians have been more successful than have data miners, and why so
- Should statistics be integrated with database managing, data mining, making predictions, and discovering and evaluating knowledge
- Favorite uses of statistics in data mining

Speakers

Michael I. Jordan, Department of Electrical Engineering and Computer Science, Department of Statistics, University of California, Berkeley

"Integration of heterogeneous data: Kernel methods and graphical models"

A pervasive problem in data mining is that of integrating data from multiple, heterogeneous sources. Data sources may have varying formats, semantics and degrees of reliability. I illustrate this problem in two areas: one involving the combination of text and images, and the other involving the prediction of the cellular location of proteins. In the case of text and images, I describe a methodology for solving data integration problems that is based on probabilistic graphical models, a formalism that exploits the conjoined talents of graph theory and probability theory to build complex models out of simpler pieces. In the case of protein annotation, I discuss an approach based on "kernel methods," an area of machine learning that makes significant use of convex optimization techniques. I show how multiple kernels can be combined, yielding a problem that is still within the scope of convex optimization. (with David Blei, Nello Cristianini, Gert Lanckriet and William Noble)

Leo Breiman, Department of Statistics, University of California, Berkeley

"Observations on Approaches to Data Mining Diversity"

Data mining has come to mean something big and amorphous and very "hollywoodish." If we think of it as the art of coaxing desired information out of data, then the approaches vary widely depending on the field, the questions, and the data. The questions and methods for the understanding of microarray data are far different from those for detecting credit card fraud or setting up the association rules for Amazon.com. In some applications, the data base is already there and the question is how to structure it to answer questions the client is interested in. In others, the question is how to gather data to resolve a significant problem. In other words, data mining refers to a huge and unstructured enterprise, in which we can best learn by looking at specific examples and searching for areas of commonality. I will discuss some of the latter.

Doug Nychka, National Center for Atmospheric Research