OPPORTUNITIES FOR THE MATHEMATICAL SCIENCES
JUNE 26-27, 2000 WORKSHOP

Statistical Issues for Databases, the Internet, and Experimental Data

M. BOCK

A. STATISTICAL ANALYSIS ISSUES FOR DATABASES IN BIOLOGY AND OTHER FIELDS

The wealth of data from research has the potential to drown the unprepared or at least to render useless much of the expertise and effort that went into developing it. The data are great in number, complex, often of extremely high dimension, and frequently noisy. The organization of the data in sophisticated databases has opened the door to its use. The storage of genome and protein data in publicly available databases significantly increases the quality of research for all in the biological sciences, and similar successes exist in other disciplines. However, the bottleneck has become the availability of good statistical techniques that quickly and effectively compare and analyze large numbers of data sets simultaneously. Statistical techniques for data mining which involve string matching, cluster analysis, and searches for underlying patterns are vital to opening the bottleneck. A critical need exists for descriptions of the statistical variability of these techniques so that users may interpret the quality of the answers they receive. Accompanying measures of the quality and variability of the data involved are also essential.

The value of these search and analysis techniques is not limited to the public databases. Results of individual experiments are often stored in private databases which become essential for documenting every step of highly complex experimental process. Gene and protein expression data from microarrays and other technologies would seem limitless as one generates it in many environments and along the full life time scale of organisms. Once stored in private or public databases, the multiple analyses are beyond the scope of an individual who may never be able to look at all the data. They benefit greatly from the computationally intensive approaches that pull out patterns for further investigation. With the growing popularity of methods and tools for data mining, it is increasingly critical that these be underpinned with solid statistical theory and understanding that will validate their usage in practice.

As databases proliferate, the problems associated with combining them arise too. Statistical techniques for deciding when to merge slightly different files from different databases become critical. The demand to merge databases is growing and provides pressure to create statistical techniques that do not allow the quality of the data to degrade.

B. STATISTICAL ISSUES FOR THE INTERNET

The growth of the Internet and electronic commerce is creating a variety of statistical challenges. What metrics should one use to gauge performance? How can one move from "pretty pictures" to visual displays that convey proper statistical impressions of the state of the network? How can local information be gathered into synthesized global views? These and other such challenges are made more complicated by the need to distinguish transient phenomena from steady-state characteristics and underlying trends.

Search engines are fundamental tools for retrieving information over the World Wide Web, and these tools are implicitly or explicitly statistical in nature. A major challenge is to improve the quality and timeliness of the information received. Statistical methods can help to define what makes documents "similar" and how to locate clusters of like documents efficiently. This includes the need for cross-language retrieval.

C. STATISTICAL ANALYSIS ISSUES FOR PHYSICS EXPERIMENTAL DATA

The high effort, cost and, sometimes, rarity of data from physics experiments make it imperative that the maximum amount of information be available. This would be enhanced by the further development and use of statistical techniques that provide reliability statements, often in the form of confidence regions for simultaneous measurements. Classical frequentist statistical interpretations of measurements give reliability statements that describe the variability of often highly precise measurements in the context of large numbers of potential independent repetitions of the experiment. However, the variability of some types of noisy measurements are best described by including prior information or beliefs about sources of variability through Bayesian analyses. Highly computationally intensive techniques have made it possible to incorporate diverse types of prior distributions for parameters of the models for the measurements and to compute the resulting reliability statements. An even greater diversity of models and priors is needed.

Last Modified:
Oct 17, 2001

Previous page | Top of this page | Next page

Policies and Important Links | Privacy | FOIA | Help | Contact NSF | Contact Web Master | SiteMap

The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: (703) 292-5111, FIRS: (800) 877-8339 | TDD: (800) 281-8749

Last Updated:
10/17/01
Text Only