Northeastern Illinois Chapter of the American Statistical Association

Volume 18, Issue 2
Summer, 2003

Statfax . . .

19th Annual Summer Workshop

Data Mining: Where Do We Go From Here?

This conference is intended both for those who wish to learn what data mining is all about and for those who have experience using data mining techniques but who would like to know how to use these techniques more effectively.

The sheer volume and complexity of data collected or available to most organizations has created an imposing barrier to its effective use. These challenges have propelled data mining to the forefront of making profitable and effective use of data. Data mining is a process that uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that may be used to make accurate predictions.

While the most widespread applications of data mining are in CRM (customer relationship management) some of the other important applications include fraud detection and identifying good credit risks.

The first and simplest analytical step in data mining is to describe the data — for example, summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look at the distribution of values of the fields in your data. But, the standard exploratory data techniques of graphing and summarizing each variable take too long when dealing with hundreds of candidate predictors. Making scatterplots of each pair is even less feasible.

But data description alone cannot provide an action plan. You must build a predictive model based on patterns determined from known results, then test that model on results outside the original sample. In classical data analysis, the exploratory phase usually precedes the model selection phase. It's seen as a necessary preliminary for understanding the data before beginning to think about how to model it. But in data mining, sometimes we start with a preliminary model just to narrow down the set of potential predictors. This exploratory data modeling (EDM) seems to be at odds with standard statistical practice, but, in fact, it's simply using models as a new exploratory tool.

In this course, we'll take a brief tour of the current state of data mining algorithms and using several case studies to explain how EDM can be used to narrow the search for a predictive model and to increase the chances of producing useful and meaningful results.

The outline for this workshop is as follows:

OVERVIEW OF DATA MINING

- Why do data mining?
- Types of models: *predictive* (classification, regression, time series); *descriptive* (clustering, association detection, sequence detection)
- The data mining process

BUILDING THE MINING DATABASE

- Stating the business problem
- Description of the data sets
- Enriching the data with external data sources

UNDERSTANDING THE DATA

- Graphical methods
- Selecting data: columns (reducing dimensionality); rows (sampling)
- Transforming the data: data representation (scaling, binning, encoding)
- Creating new attributes

BUILDING THE MODEL

- Commonly used algorithms: classical regression (linear and non-linear), logistic regression, decision trees, neural nets, K-nearest neighbor, MARS, clustering
- Algorithm characteristics
- Choosing appropriate algorithms: matching algorithms to the business problem; matching algorithms to the data; no best algorithm

THE MODEL BUILDING CYCLE

- Using models to explore
- The cycle of model building

VALIDATING THE MODEL

- Need for validation
- Simple validation

MODEL EVALUATION

- Confusion matrices
- Lift and ROI curves

WHAT CAN GO WRONG

- Overfitting
- Performance
- Interpretation
- Model limitations

Dick De Veaux
Williams College

Dick De Veaux holds degrees in Civil Engineering (B.S.E. Princeton), Mathematics (A.B.Princeton), Physical Education (M.A. Stanford; Specialization in Dance) and Statistics (Ph.D., Stanford). He has taught at the Wharton School, the Princeton University School of Engineering, and, since 1994, has been a professor in the Math and Stat Department of Williams College. He has won numerous teaching awards including a "Lifetime Award for Dedication and Excellence in Teaching" from the Engineering Council at Princeton. He has won both the Wilcoxon and Shewell awards (twice) from the American Society for Quality and was elected fellow of the ASA in 1998. He has served as General Methodology Chair for the JSM Program Committee 3 times, in 1987, 1995 and 1999. He served as program chair for SPES in 1996. He was the Program Chair for the 2001 JSM in Atlanta.

Dick has been a consultant for nearly 20 years for such companies as Hewlett-Packard, Alcoa, First USA Bank, Dupont, Pillsbury, Rohm and Haas, Ernst and Young, General Electric, and Chemical Bank. He holds two U.S. patents and is the author of over 25 refereed journal articles. His hobbies include cycling, swimming, singing (barbershop, doo wop and classical -- he is the head of the Diminished Faculty, a local doo wop group) and dancing (he was once a professional dancer and has a masters degree in dance education). He is the father of four children ages 6, 8, 10 and 12. He is currently on sabbatical at the University Paul Sabatier in Toulouse France and is the co-author, with Paul Velleman, of an introductory textbook titled "Intro Stats" to be published by Addison-Wesley in May 2003.

~Bala Hosmane

Summer Workshop

The summer workshop will be held on June 12, 2002, from 8:30am to 4:30pm, at the Adam's Mark Hotel, 2875 N. Milwaukee Ave., Northbrook, IL 60062. (Telephone: 847-298-2525.) The luncheon menu will be finalized after the newsletter is published. If you have special needs, please contact Renee Alpern at (708-202-4897) or email: alpern@research.hines.med.va.gov.

~Renee Alpern

Program Committee

I would like to take this opportunity to thank everyone that participated in the Spring Chapter Meeting. I would especially like to thank the three speakers, Daniel Frobish, Kiang Liu, and

Art Roth, for helping to make the Spring Chapter Meeting a success.

Next up is the Summer Workshop, which is scheduled for June 12th. Professor Dick De Veaux from Williams College has graciously agreed to present a workshop entitled Data Mining: Where do we go from here? I would like to thank Jay Chmiel for suggesting the topic for the Summer Workshop, and for convincing Dick to participate.

The speaker line-up for Fall Chapter Meeting is complete, with Sanjib Basu, Brent Logan, and Domenic Reda scheduled to participate. The Fall Chapter Meeting has been scheduled for October 23, 2003. Additional details will be provided later this summer.

~Bala Hosmane

Treasurer's Report

BALANCE 04/01/2003	**4,694.13**
INCOME	
2003 Spring Meeting	2,644.00
Bank Interest (thru 3/31/2003)	13.98
Total Income	2,657.98
TOTAL ASSETS	7,622.11
EXPENSES	
2003 Spring Meeting	(2,286.90)
Design Conf. Support	(500.00)
Bank Charges	(6.00)
Total Expenses	(2,792.90)
BALANCE 12/31/2002	**4,829.21**

2002 Paid Memberships	138
Non-Students	130
Students	8

~ Harold Frush

Chapter Website

Don't forget to visit our Chapter web page at http://www.amstat.org/chapters/NortheasternIllinois. The website can also be accessed by following links found on the main ASA web page (www.amstat.org). Select *Chapters*, then *Chapter Links by District*, then *North Eastern Illinois* (Under District 4), and finally, *visit the Northeastern Illinois Chapter home page*.

The Chapter website has lots of information about our Chapter, including events, program information, a listing of members, and contact information for Chapter officers and committee members. You can also peruse PDF versions of current and past Chapter newsletters. Finally, if the speaker provides them, slides/notes from previous presentations are available to view or download.

~Paul Cernohous

Calendar

The following are some upcoming regional and national events that may be of interest to you. Additional details can be found at the ASA website (www.amstat.org/dateline).

- Conference on New Directions in Experimental Design (May 14-17, 2003; Chicago, Illinois)
- 26th Annual Midwest Biopharmaceutical Statistics Workshop (May 19-21, 2003; Muncie, Indiana)
- Spring Research Conference on Statistics in Industry and Technology (June 04-06, 2003; Dayton, Ohio)
- Graybill Conference – Microarrays, Bioinformatics, and Related Topics (June 18-20, 2003; Fort Collins, Colorado)
- Joint Statistical Meetings (August 03-07, 2003; San Francisco, California)
- FDA/Industry Workshop (September 18-19, 2003; Bethesda, Maryland)

In addition, "CART, MARS, and MART/Treenet Training Seminars" will be presented in Chicago, Illinois (May 19-22, 2003). For additional details, go to http://www.salford-systems.com/training.html. If you decide to attend one of these training sessions, you may be entitled to a Chapter discount.

Officers and Committee Chairs

PRESIDENT
Richard Rode
(847) 937-3757
Richard.A.Rode@abbott.com

PRESIDENT-ELECT AND PROGRAM CHAIR
Bala Hosmane
(815) 753-6858
bala@math.niu.edu

SECRETARY
Theodora Cohen
(847) 810-4302
tcohen@neophrm.com

TREASURER
Harold Frush
(847) 937-5332
harold.frush@abbott.com

PAST-PRESIDENT
Domenic Reda
(708) 202-5853
reda@research.hines.med.va.gov

CHAPTER REPRESENTATIVE
Art Roth
(847) 982-4775
arthur.j.roth@pharmacia.com

ARRANGEMENTS
Renee Alpern
(708) 202-4897
alpern@research.hines.med.va.gov

Peter Yu
(847) 236-2359
peter.yu@tappharma.com

MEMBERSHIP CHAIR
Jaime Delgado
847-982-7985
Jaime.Delgado@Pharmacia.com

AWARDS AND RECOGNITION
Prof. A. S. Hedayat
(312) 996-4831
hedayat@uic.edu

COMMUNITY ACTIVITIES
Carole Bernett
(630) 773-2697
cbernett@ix.netcom.com

Paul Coe
(708) 524-6640
coepaul@email.dom.edu

NEWSLETTER EDITOR
Anita Ross
(847) 675-7807
aross@northpark.edu

WEBMASTER
Paul Cernohous
(847) 938-0316
paul.cernohous@abbott.com

Statfax is published three times a year (Spring, Summer, and Fall), by the Northeastern Illinois Chapter of the American Statistical Association. **Statfax** welcomes letters and material for articles from its readers. Address correspondence to: The Editor, Statfax, c/o Anita Ross, Ross Consulting, 4554 Main St., Skokie, IL, 60076.

The Chapter's purpose, as stated in its Constitution, is "to foster statistics and its applications and promote the interests of the statistical profession…". Accordingly, diverse views are presented. They do not necessarily reflect the opinions of the officers or the policies of the Chapter.