Data Mining: A Long-Term Dream

David Waltz, NEC Research Institute
Se June Hong, IBM T.J. Watson Research Center

In 1974, Marvin Denicoff of ONR (Office of Naval Research) told me [David Waltz] about a problem that the US Navy had. One of their destroyers had suffered an explosion in its boiler, which killed one or more sailors. A team of Navy investigators found that the boiler had been the source of many problems and repairs over the years and that the data had been encoded and saved in a database. However, no process existed for identifying patterns of problems, so the data simply lay unused and ignored, and the underlying problems with the boiler went unrecognized. The admiral in charge blew his top, and Marvin funded a sizable program (from which I received my first research funds as a new faculty member at the University of Illinois) to see that this kind of problem would eventually be discovered automatically. One of my students, Doug Dankel (now at the University of Florida), did his thesis on the mining of this data (we didn't actually use the term data mining at that time). Doug's program constructed standard models (for example, Gaussian distributions) for variables and selected outliers, and then looked for records and entities that had conjunctive outliers, and so on. (Alas, he never published this except as, perhaps, a conference paper, despite my nagging—I thought it was a great idea, similar to that used later by Doug Lenat in his AM system to find interesting mathematical concepts.[1,2])

—David Waltz

1094-7167/99/$10.00 © 1999 IEEE

THE MORAL OF THE PREVIOUS anecdote still holds true today: we are much better at collecting data than we are at using it in a sensible way, and the amounts we collect outstrip our ability to use them with existing methods. This issue offers 10 fine examples of ways to extract valuable information from our mountains of data along with visions of how to ultimately redress the balance between collecting and understanding data.

What's in this issue

Although the term data mining became popular much less than a decade ago, it has become an important field attracting attention from both industrial users and R&D workers. Gregory Piatetsky-Shapiro describes the growing community of data mining in "Expert Opinion." Data mining is becoming an indispensable decision-making tool in the ever more competitive business world. And challenging applications inspire new techniques and affirm their utility.

Two invited articles by Simon Kasif and Steven Salzberg discuss data mining in the exciting field of computational biology, where frequent pattern discovery, clustering, and classification all play crucial roles in understanding protein structures and their functions.

The next group of articles represents innovative data-mining applications in more traditional domains with conceptually flat data tables. Chidanand Apte, Edna Grossman, Edwin Pednault, Barry Rosen, Fateh Tipu, and Brian White have developed a specialized probabilistic model for auto insurance pure premium—that is, the expected claim amount for each policy holder that meets the strict actuarial requirement. Special attention is given to missing values and the model's scalability. Sylvain Létourneau, Fazel Famili, and Stan Matwin describe the entire process of modeling failing aircraft components, from gathering data and generating serious models to evaluating them using a domain-specific scoring function. Extensive experiments seem to show that the nearest-neighbor method does best for most parts. Philip Chan, Wei Fan, Andreas Prodromidis, and Salvatore Stolfo apply a specialized boosting technique to credit card fraud detection for scalability and enhanced utility of the model. Cost savings thus achieved demonstrate the improved accuracy of multiple models while providing a scalable fast model generation on a very large data set.

Due to *Intelligent System*'s page constraints, the next four articles of this special issue will appear in a follow-up issue in March 2000. These articles will deal with mining nontraditional forms of data. Two articles deal with text mining. Sholom Weiss, Brian White, Chidanand Apte, and Fredrick Damerau show that simple and fast document matching can effectively assist help-desk applications in matching problem descriptions to relevant stored solution descriptions. Kurt Bollaker, Steve Lawrence, and C. Lee Giles present their Web-based CiteSeer system that finds user-specific scientific documents in the documents that are preprocessed and stored in a local database. The system learns the user profile from the interactions and uses it as an agent to monitor new documents that might interest the user.

Two articles describe new techniques. Neal Lesh, Mohammed Zaki, and Mitsunori Ogihara find frequent subsequence patterns that help classify a sequence, such as DNA, into different classes of sequences. Their FeatureMine system efficiently examines subsequences to select a drastically pruned feature set. The article by Diane Cook and Lawrence Holder deals with data mining graph-structured data, such as CAD diagrams and chemical structures. Their Subdue system uses a beam search to discover substructures (features) that efficiently describe the given set of structures based on the MDL principle. Both of these new techniques discover useful features from the source data that are not the flat tables typically found in popular databases, which discriminate or describe such data.

Considering the increasing need for discovering useful relations and information from nontraditional database data, we feel that data mining ought not be "just another aspect of database systems," begging to differ from Gregory Piatetsky-Shapiro's last observation in his article.

The future of data mining

As mentioned earlier, we are much better at collecting data than we are at using it. However, using data sensibly requires superhuman AI—that is, the ability to do intelligent things that people can't because of the vast amount of computation needed. For this reason, this is an area where clever and creative work on the part of a machine is possible—and needed—for good solutions. Today, much data mining is applied statistics, but, eventually, we should use the models and knowledge we formulate to locate proverbial needles in haystacks or to find patterns in vast search spaces. Raw, brute-force search, although useful, will only get us so far. For example, we might need planning to select what data to consider before doing a detailed analysis.[3] Such bootstrapping of data mining by use of knowledge is just beginning. Eventually, knowledge discovery and more systematic knowledge acquisition through machine learning could be data mining's holy grail, as tomorrow's data will come in much more diverse forms than what we typically find today. ■

References

1. D.B. Lenat, "The Ubiquity of Discovery," *Artificial Intelligence*, Vol. 9, No. 3, Dec. 1977, pp. 257–285.
2. D.B. Lenat, "On Automated Scientific Theory Formation: A Case Study Using the AM Program," *Machine Intelligence 9*, 1979, J. Hayes, D. Michie, and L.I. Mikulich, eds., Halstead, New York, pp. 251–283.
3. S. Chien et al., "Using Artificial Intelligence Planning to Automate Image Data Analysis," *Intelligent Data Analysis*, Vol. 3, No. 3, Aug. 1999, pp. 159–176.

David Waltz has been vice president of the NEC Research Institute's Computer Science Research Division and adjunct professor of computer science at Brandeis University since 1993, and will become president of the NEC Research Institute in April 2000. His research interests have included constraint propagation, computer vision, massively parallel systems for both relational and text databases, memory-based and case-based reasoning systems and their applications, protein-structure prediction using hybrid neural-net and memory-based methods, and connectionist models for natural-language processing. He received his SB, SM, and PhD, all in electrical engineering from MIT. He is past president of the AAAI, a fellow of the ACM and AAAI, a senior member of the IEEE, and a former chairman of ACM SIGART. Contact him at NEC Research Institute, 4 Independence Way, Princeton, NJ 08540; waltz@research.nj.nec.com.

Se June Hong is a research staff member working on data-mining technology at IBM T.J. Watson Research Center in Yorktown Heights, New York. His research interests have included error-correcting code, fault-tolerant computing, design automation, and knowledge-based systems. He received his BSc degree in electronic engineering from Seoul National University and his MS and PhD in electrical engineering from the University of Illinois. He is a fellow of the IEEE; a member of the ACM, AAAI, KSEA, and Sigma Xi; and a foreign member of the National Academy of Engineering of Korea. Contact him at IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598; hong@us.ibm.com.