from *SIAM News*, Volume 32, Number 10

Data Mining Emerges as A New Discipline in a World of Increasingly Massive Data Sets

By James Case

On May 12, the day shared by the Sixth SIAM Conference on Optimization and the 1999 SIAM Annual Meeting in Atlanta, Olvi Mangasarian of the University of Wisconsin gave a joint invited talk, "Optimization in Machine Learning and Data Mining."

Unlike machine learning, which has occupied the AI community for years, data mining is a discipline of relatively recent origin. The term refers to a collection of techniques for extracting useful information from large data sets—too large, in many cases, even to be loaded into main memory. In a world where even the simplest daily task—like making a phone call, using a credit card, or buying hardware and groceries—leaves an electronic footprint, an increasing number of increasingly important data sets fit that description.

Automated data collection devices, capable of generating terabytes (a.k.a. terrorbytes) and even petabytes of information at rates measured in gigabytes per hour, are rendering existing inference methods obsolete. The biggest data warehouse in the world—the Wal-Mart system, built on an NCR platform called Teradata—reportedly contained 11 terabytes (11×10^{12} bytes) of information as of 1996 and has doubtless continued to expand. The satellites of NASA's Earth Observing System are capable of generating more than a terabyte of data per day. In such a world, even the simplest browsing operation can result in an avalanche of useless and irrelevant data.

Imagine a program for deciding whether two rows of data differ in more than "a few fields." While such "find similar" problems appear simple and can be approached in various ways, executing any one of them on a massive data store is by no means trivial. More complicated questions could require the analysis of millions of data points residing in a space of a thousand dimensions. Who can do that?

Olvi Mangasarian, whose long-dormant early work on pattern recognition found a genuine "killer ap"—breast cancer diagnosis—as a result of a chance dinner-party conversation.

From Raw Data to "Documented Knowledge"

Among the more predictable results of the emerging discipline has been a new wave of multiletter acronyms. Perhaps the best known are OLAP, for On-Line Analytical Processing, and KDD, for Knowledge Discovery in Databases. As used at Microsoft, the former term refers to a database capable of responding to queries more complex than those handled by the standard "relational" databases of the 1970s and 1980s, while the latter refers to a newer and even more versatile generation of software. OLAP has been prolific in its output of re-lated acronyms, including MOLAP (multi-dimensional OLAP), ROLAP (relational OLAP), HOLAP (hybrid OLAP), and most recently DOLAP (desktop OLAP).

OLAP systems rely heavily on precomputed aggregates—obtained after a single pass through the data, by summing or averaging over particular indices and groups of indices—as well as projections onto lower-dimensional subspaces of the high-dimensional space in which raw data so often reside. Because the number of potential aggregates increases exponentially with the number of dimensions, much of the work in OLAP systems involves deciding which aggregates to precompute, and how to de-rive (or estimate) additional aggregates from those that have been precomputed.

OLAP exploration is guided by user-supplied instructions regarding the histograms to be created, the variables to be plotted against one another, and the level of detail employed. Inference and modeling are left to the user, who is expected to recognize patterns of interest via visualization in lower-dimensional subspaces, and to formulate testable hypotheses concerning the reduced data sets furnished by the system. KDD, in contrast, combines methods from database theory, statistics, pattern recognition, AI, high-performance computing, and the like, in an effort to discover patterns hidden within the data and model the behavior that produced them.

A pattern might be a simple data summary, a data segmentation, or a model of dependencies (a.k.a. links) within the data. KDD, intended to lead all the way from raw data to "documented knowledge," proceeds in several steps. Because the raw data generated by industrial processes like manufacturing, telephone switching, and customer billing are often recorded at remote locations in arcane formats, KDD cannot even begin until all have been assembled, cleaned up, and organized into what is called a "data warehouse." From that, a subset of the data that is relevant to a given project must be extracted—rather as a student might borrow relevant books from a library before writing a term paper—followed by the formation of natural aggregates between which causal relationships can be expected to exist.

The mere construction of a data warehouse is often enlightening. One firm, for instance, reportedly discovered that its records

contained 27 separate and distinct spellings of the name K-Mart. Others have uncovered missing fields, conflicting reports, and all manner of other debilitating flaws.

Even if each step of the KDD process "runs" successfully, the results will not necessarily be informative. A group at Silicon Graphics was amused to learn—from an early version of SGI's own data mining tool MineSet—that "with 99.7% certainty, all individuals who are husbands are also males." While obvious to humans, such conclusions are every bit as intriguing to a computer as patterns that flag double billing by physicians in Australia, an incipient epidemic in Maine, or a "softening" of U.S. demand for Sport Utility Vehicles.

A famous result of first-generation KDD analysis was the observation that the most frequent late-night purchases in supermarkets are diapers and beer. As a result, strategically located aisles are frequently blocked after 10 PM to route diaper purchasers past the beer cooler on their way to checkout. A new generation of data mining tools now emerging, although not yet simple enough to accommodate corporate end-users, are able to detect patterns of the foregoing sort while being significantly easier to use than the options previously available as add-ons to traditional statistical analysis packages.

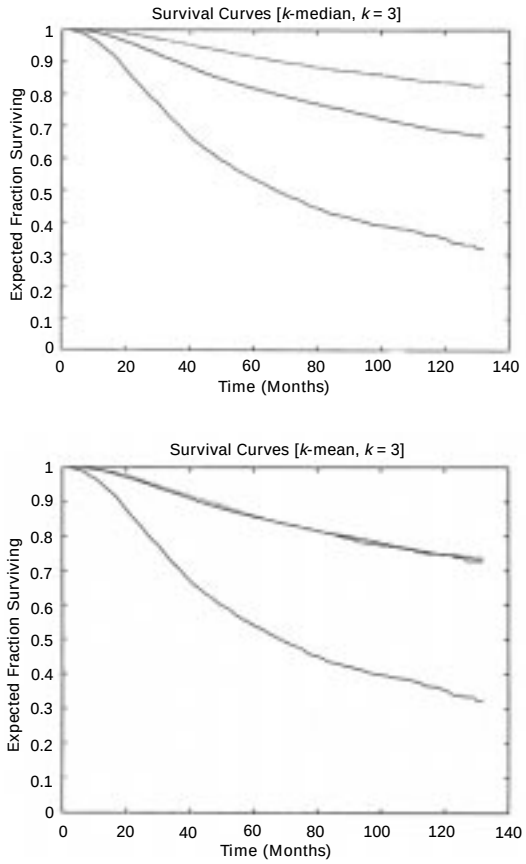

Figure 1. *Visual evidence that k-medians (top) can be more useful than k-means (bottom) for practical purposes. The rival methods were used to separate 21,960 data points from a National Cancer Institute database into three clusters; survival frequencies for each cluster were then plotted against time.*

Applications

The most promising applications of KDD are in fields requiring high-payoff knowledge-based decisions in rapidly changing and information-rich environments. Perhaps the most persuasive business application is "database (i.e., mail-order) marketing," which relies on customer information to tailor offers to particular segments of the market. Affluent suburban customers with expensive tastes, for example, will receive the "sneak preview issue" of ABC Outfitters' autumn catalogue some weeks before those with less promising credentials and/or purchase records receive the ordinary issue, and months before proven misers receive the "remainders" issue. Needless to say, the prices decline from issue to issue.

Because scientific users typically know their data in intimate detail, it may be easier to develop KDD applications in science than in retailing, finance, or other areas of commerce. A case in point is the 2nd Palomar Observatory Sky Survey, which took more than six years to complete and gathered more than 3 terabytes of image data, in which an estimated 2 billion sky objects are "visible." The 3000 photographic images were scanned into digital format, with 23,040 × 23,040 pixels per image and a resolution of 16 bits per pixel.

An automatic method was needed for identifying and then classifying (as star, galaxy, quasar, black-hole accretion disc, and so forth) the many objects in each digital image. The majority of objects are faint, visible to the naked eye but impossible to categorize by visual means. To meet this need, a team from the Jet Propulsion Laboratory developed the Sky Image Cataloging and Analysis Tool (SKI-CAT). Once the basic image segmentation was complete, 40 attributes deemed important by astronomers were measured for each object identified. SKI-CAT then discarded 32 of the 40 measurements and devised a classification scheme exploiting the remaining eight.

To test SKI-CAT's accuracy, a small sample of the classified objects were re-classified by a far more expensive method. The two classifications agreed on 94% of the sample. SKI-CAT subsequently helped astronomers to discover—in record time—16 new high-red-shift quasars. Such objects, among the most distant (and therefore oldest) in the universe, are extremely difficult to find. But they provide rare and valuable clues about the early history of the universe.

KDD methods have also been used to count and locate the mountains on Venus (from the Side Aperture Radar data obtained by the Magellan spacecraft during its five years in orbit around that most Earth-like of planets); to identify individual genes in the human genome; to detect and measure tectonic activity in the Earth's crust (from satellite data); and to estimate the duration and strength of tropical cyclones (from massive amounts of electronically gathered atmospheric data). As more sensitive data collection methods become available along these and other avenues of inquiry, ever more sophisticated methods of analysis will be called for.

Computational Considerations

The focus in Mangasarian's Atlanta talk was on computational methods. In his first two papers on pattern recog-nition (1965 and 1968), he told the audience, he had proposed the repeated use of a linear programming routine, called as a subroutine by a main program designed for a more specific purpose. It was a novel suggestion at the time. Underwhelmed by the response, he shifted his focus to other matters. Another twenty years were to pass before a chance dinner-party conversation led him to a genuine "killer ap" for his long-dormant work on pattern recognition. His description of both the application and the method appeared in the

September 1990 issue of *SIAM News*.

The original problem Man-gasarian considered was the separation of two finite and disjoint sets of points in R^n—think of them as *x*'s and *o*'s drawn on a blackboard of arbitrary (but again finite) dimension—by means of piecewise-linear discriminant functions. If the convex hulls of the two sets are disjoint, the problem is trivial: A single hyperplane separates the convex hull of the *x*'s from that of the *o*'s, and (signed) distance from that hyperplane serves as a (linear) discriminant function. Moreover, solution of a single linear program suffices to determine the separating hyperplane.

If the convex hulls intersect, however, no single hyperplane suffices. In that case, Man-gasarian demonstrated (in a 1965 paper) the utility of a method that finds the thinnest closed hyperslab H_1 (a subset of R^n bounded by parallel hyperplanes) for which (a) the intersection of the two convex hulls is contained in H_1 and (b) each of the open halfspaces complementary to H_1 contains only *x*'s or only *o*'s. H_1 can also be determined via the solution of a single linear program. The *x*'s and *o*'s exterior to H_1 are then regarded as "already separated," and a hyperplane separating only the remaining *x*'s and *o*'s is needed. If one is found, the process stops with all the *x*'s separated from all the *o*'s by a piecewise-linear discriminant function; otherwise, the process continues.

Eventually, the process terminates with the *x*'s and *o*'s separated by a piecewise-linear discriminant function. (The reader may find it instructive to construct the required function for three *x*'s and three *o*'s situated at alternate vertices of a regular hexagon.) In rare degenerate cases, the exterior of the most recently computed hyperslab contains neither *x*'s nor *o*'s, necessitating the insertion of an anti-degeneracy step between successive linear programs.

Even in the presence of degeneracy, the process can always be carried to completion, given only that no point of R^n is both an *x* and an *o*. It then makes sense to infer that any uncategorized point for which the discriminant function assumes a positive value is another *x*, while those for which the discriminant function assumes negative values are *o*'s. If consideration of additional *x*'s and *o*'s shows that assumption to be erroneous, the larger sets of *x*'s and *o*'s can be used to "retrain" the classification scheme. More recently, it has been demon-strated that Mangasarian's "hyperslab method" constitutes an effective method for training neural networks with partially preassigned weights.

Beyond the Killer Ap

The killer ap brought to Mangasarian's attention at the fateful dinner party was breast cancer diagnosis. Other applications have continued to surface, and he alluded—albeit telegraphically—to a number of them in his Atlanta talk. A particularly interesting one concerned the authorship of the disputed Federalist papers. The original method has been generalized in numerous directions, including separation into more than two sets, separation by nonlinear hypersurfaces, and the classification of as many as a million points in R^{32}. As formulated by Mangasarian, all such classification problems can be reduced to the minimization of concave (piecewise-linear) objective functions subject to (numerous) linear constraints.

Such problems, relatively routine in spaces of low dimension, become unwieldy when the constraint matrix exceeds main memory capacity. To avoid that difficulty, he and his co-workers have developed a "chunking" method for solving such problems subject to a small, randomly chosen subset of the given constraints. The constraints found to be inactive are discarded, and the problem is solved again subject to the union of all currently active constraints with another small random subset of the given ones; these steps are repeated un-til convergence. Convergence was achieved after 25.9 hours for a problem with 5×10^5 constraints, and after 231.3 hours for another involving 10^6 constraints.

Two other important problems in data mining, the so-called *k*-mean and *k*-median clustering problems, constitute alternative ways of identifying the most densely populated subregion of the region in which given data points reside, along with the second most densely populated subregion, . . . , up to and including the *k*th most densely populated subregion. The *k*-median problem, Mangasarian pointed out, can also be solved by minimization of a piecewise-linear concave objective function subject to linear constraints.

He produced visual evidence that *k*-medians are sometimes more useful than *k*-means for practical purposes: Using the rival methods, he separated 21,960 data points in a National Cancer Institute database into three clusters and plotted survival frequencies for each cluster against time (Figure 1). The survival characteristics of the three groups identified by the *k*-median method ($k = 3$) differ markedly, while those of two of the three groups identified by the *k*-means algorithm do not. Prospects for the high-risk groups identified by the two methods, on the other hand, are virtually identical.

It was evident to all in the audience that much has been accomplished in this important emerging field, that much remains to be done, and that Mangasarian, his students, and his various other co-workers are leaders in the field.

James Case is an independent consultant who lives in Baltimore, Maryland.