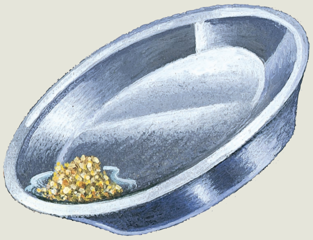

Mining Scientific Data

Usama Fayyad, David Haussler, and Paul Stolorz

Digesting millions of data points, each with tens or hundreds of measurements—generally beyond a scientist's human capability—can be turned over to data mining techniques for data reduction, which functions as an interface between the scientist and large datasets.

THE SCIENTIST AT THE OTHER END OF today's data collection machinery—whether a satellite collecting data from a remote sensing platform, a telescope scanning the skies, or a microscope probing the minute details of a cell—is typically faced with the problem: What do I do with all the data? Scientific instruments can easily generate terabytes and petabytes of data at rates as high as gigabytes per hour. There is a rapidly widening gap between data collection capabilities and the ability to analyze the data. The traditional approach of a lone investigator staring at raw data in pursuit of (often hypothesized) phenomena or underlying structure is quickly becoming infeasible. The root of the problem is that data size and dimensionality are too large. A scientist can work effectively with a few thousand observations, each having a small number of measurements, say five. Effectively digesting millions of data points, each with tens or hundreds of measurements, is another matter.

When a problem is fully understood and the scientist knows what to look for in the data through well-defined procedures, data volume can be handled effectively through data reduction.[1] By reducing data, a scientist is

[1] Data reduction is a term used in science data analysis to refer to the extraction of essential variables of interest from raw observations. Particularly appropriate when dealing with image datasets, it involves transformation, selection, and normalization operations.

TERRY WIDENER

effectively bringing data size down to a range that is analyzable.

In scientific investigation, because we are often interested in new knowledge, effective data manipulation and exploratory data analysis looms as one of the biggest hurdles in the way of exploiting the data. In this article, we give an overview of the main issues in the exploitation of scientific datasets through automated methods, present five case studies in which knowledge discovery in databases (KDD) tools play important and enabling roles, and conclude with future challenges for data mining and KDD techniques in science data analysis.

Data Reduction and Data Types

DATA mining and KDD techniques for automated data analysis can and do play an important role as an interface between scientists and large datasets. Machines are still far from approaching human abilities in the areas of synthesis of new knowledge, hypothesis formation, and creative modeling. The processes of drawing insight and conducting investigative analyses are still clearly in the realm of tasks best left to humans. However, automating the data reduction procedure is a significant niche suitable for computers. Data reduction involves cataloging, classification, segmentation, partitioning of data, and more. It is the most tedious stage of analysis, typically involving manipulation of enormous amounts of data. Once a dataset is reduced (say to a catalog or other appropriate form), the scientist can proceed to analyze it using more traditional (manual), statistical, or visualization techniques. The higher levels of analysis include theory formation, hypothesis of new laws and phenomena, filtering what is useful from background, and searching for hypotheses that require a large amount of highly specialized domain knowledge.

Machines are still far from approaching human abilities in the areas of synthesis of new knowledge, hypothesis formation, and creative modeling.

Data comes in many forms—from measurements in flat files to mixed (e.g., multispectral/multimodal) data including time series (e.g., sonar signatures and DNA sequences), images, and structured attributes. Most data mining algorithms in statistics and KDD [3] (see also Glymour's article in this special section) are designed to work with data in flat files of feature vectors.

Data types include:

Image data. Common in science applications, image data offers unique advantages in that it is relatively easy for humans to explore and digest. On the other hand, image data poses serious challenges on the data mining side. Feature extraction is the dominant problem; using individual pixels as features is typically problematic, since a small portion of an image easily turns into a high-dimensional vector.[2]

Time-series and sequence data. Challenges here include extracting stationary characteristics of an entire series, whether or not it is stationary; if it is not stationary (e.g., in the case of DNA sequences), segmentation is needed to identify and extract nonstationary behavior and transitions between quantitatively

[2] A small window of 30×30 pixels from an image represents a vector of 900 values. An image of 1K×1K pixels contains a very large number of these vectors (up to 106 of them).

and qualitatively different regimes in the series. An effective means for dealing with sequence data is to infer transition probabilities between process state variables from the observed data. A particularly successful class of techniques used in this type of mining is hidden Markov models (HMMs) [8], which have been extensively developed in the context of speech recognition. An HMM describes a series of observations by a "hidden'' stochastic process—a Markov process.

In the case of speech, the observations are sounds forming words, and a model represents a hidden random process that generates certain sequences of sounds, constituting variant pronunciations of a single word, with high probability. In modeling proteins, a word corresponds to a protein sequence, and a family of proteins with similar structure or function can be viewed as a set of variant pronunciations of a word. This observation allows a large amount of mathematical and algorithmic HMM machinery developed in the context of speech processing to be adapted and applied to protein modeling, greatly reducing implementation and development time and allowing impressive results to be obtained quickly [5].

Numerical measurements vs. categorical values. While a majority of measurements (e.g., pixels and sensors) are numeric, some notable examples (e.g., protein sequences) consist of categorical measurements. The advantage of dealing with numerical data is that the notion of distance between any two data points (feature vectors) is easier than defining distance metrics over categorical-value variables. Many classification and clustering algorithms rely fundamentally on the existence of a metric distance and the ability to define means and centroids.

Structured and sparse data. In some problems, variables may have some structure to them (e.g., hierarchical attributes or conditional variables that have different meanings under different circumstances). In other cases, different variables are measured for different observations, rendering flat-file representation inappropriate.

Reliability of data (sensor vs. model data). Raw sensor-derived data is often assimilated to provide a smooth homogeneous data product. For example, regular gridded data is often required in climate studies, even when data points are collected haphazardly, raising the question of data reliability; some data points need to be dealt with especially carefully, as they may not correspond to direct sensor-derived information.

Case Studies

Five case studies illustrate the contribution and potential of KDD for science data analysis. For each case, our focus is primarily the application's impact, the reasons why KDD systems succeeded, the limitations of techniques, and future challenges.

Sky Survey Cataloging

The 2nd Palomar Observatory Sky Survey (POSS-II) took more than six years to complete. The survey consisted of 3TB of image data containing an estimated 2 billion sky objects. The 3,000 photographic images are scanned into 16-bit/pixel-resolution digital images at 23,040×23,040 pixels per image. The basic problem is to generate a survey catalog recording the attributes of each object along with its class (e.g., star or galaxy). The attributes are defined by the astronomers.

Once basic image segmentation is performed, 40 attributes per object are measured. The problem is identifying the class of each object. Once the class is known, astronomers can conduct all sorts of scientific analyses, like probing galactic structure from star and galaxy counts, modeling evolution of galaxies, and studying the formation of large structure in the universe [13]. To achieve these goals, we developed the Sky Image Cataloging and Analysis Tool (SKICAT) system [12].

DETERMINING the classes for faint objects in the survey is a difficult problem. The majority of objects in each image are faint objects whose class cannot be determined by visual inspection or classical computational approaches in astronomy. Our goal was to classify objects at least one isophotal magnitude fainter than objects classified in previous comparable surveys. We tackled the problem using decision-tree learning algorithms (see chapter 19 in [3]) to accurately predict the classes of objects. The accuracy of the procedure was verified through a very limited set of high-resolution charged-couple device (CCD) images as ground truth.

By extracting rules via statistical optimization over multiple trees (see chapter 19 in [3]), we achieved 94% accuracy in predicting sky object classes. Reliable classification of faint objects increased the number of objects classified (usable for analysis) by 300%. Hence, astronomers could extract much more out of the data in terms of new scientific results [12].

SKICAT's classification scheme recently helped a team of astronomers discover 16 new high red-shift quasars in at least one order of magnitude less observation time [4]. These objects are extremely difficult to find and are some of the farthest (hence oldest) objects in the universe. They provide valuable and rare clues about the early history of the universe.

SKICAT was successful for several reasons:

- The astronomers solved the feature extraction problem—the proper transformation from pixel space to feature space. This transformation implicitly encodes a significant amount of prior knowledge.
- Within the 40-dimensional feature space, at least eight dimensions are needed for accurate classification. Hence, it was difficult for humans to discover which eight of the 40 to use, let alone how to use them in classification. Data mining methods contributed by solving the classification problem.
- Manual approaches to classification were simply not feasible. Astronomers needed an automated classifier to make the most of the data.
- Decision-tree methods, although involving blind greedy search (see Fayyad's overview article on the KDD process in this special section) proved to be an effective tool for finding the important dimensions for this problem.

Directions being pursued now involve clustering the data. Unusual or unexpected clusters in the data might be indicative of new phenomena, perhaps even a new discovery. A difficulty here is that new classes are likely to be rare in the data (one per million observations), so algorithms need to be tuned to looking for small interesting clusters rather than ignoring them as noise or outliers.

KDD applications in science may generally be easier than applications in business, finance, or other areas—mainly because science users typically know their data in intimate detail.

Finding Volcanoes on Venus

The Magellan spacecraft orbited the planet Venus for more than five years and used synthetic aperture radar (SAR) to map the surface of the planet, penetrating the gas and cloud cover that permanently obscures the surface in the optical range. The resulting dataset is a unique high-resolution global map of an entire planet. We have more of the planet Venus mapped at the 75-m/pixel resolution than we do of the Earth's surface (since most of the Earth's surface is covered by water). This dataset is uniquely valuable because of its completeness and because Venus is the most similar planet to Earth in size. Learning about the geological evolution of Venus could offer valuable lessons about Earth.

The sheer size of the dataset prevents planetary geologists from effectively exploiting its content. The first pass of Venus using the left-looking radar yielded more than 30,000 images at 1,000×1,000 pixels each. To help a group of geologists at Brown University analyze this dataset, the Jet Propulsion Laboratory developed the Adaptive Recognition Tool (JARtool) [1]. The system seeks to automate the search for an important feature on the planet—small volcanoes—by training the system via examples. The geologists would label volcanoes on a few (say 30 to 40) images, and the system would then automatically construct a classifier that would proceed to scan the rest of the image database and attempt to locate and measure the planet's estimated 1 million small volcanoes. Note the wide gap between the raw collected data (pixels) and the level at which scientists operate (catalogs of objects). In this case, unlike the case with SKICAT, the mapping from pixels to features would have to be done by the system. Hence, little prior knowledge is provided to the data mining system.

JARtool uses an approach based on matched filtering for focus of attention (triggering on candidates that vaguely resemble volcanoes and having a high false detection rate) followed by feature extraction based on projecting the data onto the dominant eigenvectors in the training data, and then by classification learning to distinguish true detections from false alarms. The tool matches scientist performance for certain classes of volcanoes (e.g., high-probability volcanoes vs. those scientists are not sure about) [1]. Limitations include sensitivity to variances in illumination, scale, and rotation. This approach does not, however, generalize well to a wider variety of volcanoes.

The use of data mining methods here was motivated by several factors:

- Scientists did not know much about image processing or about the SAR properties. Hence, they could easily label images but could not design recognizers.
- As is often the case with cataloging tasks, there is little variation in illumination and orientation of objects of interest, making mapping from pixels to features an easier problem.
- The geologists were motivated to work with us; they lacked other easy means for finding small volcanoes.
- The result is to extract valuable data from an extensive dataset. Also, the adaptive approach (training by example) is flexible and would in principle lends itself to reuse in other tasks.

DUE to the proliferation of image databases and digital libraries, data mining systems capable of searching for content are becoming a necessity. In dealing with images, the train-by-example approach, or querying for "things that look like this," is a natural interface, since humans can visually recognize items of interest, but translating those visual intuitions into pixel-level algorithmic constraints is difficult to do. Work is proceeding to extend JARtool to other applications, like classification and cataloging of sunspots.

Biosequence Databases

In its simplest computer form, the human genome is a string of about 3 billion letters containing instances of four letters—A, C, G, and T, representing the four nucleic acids, the constituents of DNA, strung together to make the chromosomes in our cells. These chromosomes contain our genetic heritage, a blueprint for a human being. A large international effort is under way to obtain this string, but obtaining it is not enough; the string has to be interpreted. DNA is first transcribed into RNA and then translated in turn from RNA into proteins to form the actual building blocks (chromosomes) of our makeup. The proteins do most of the work within the cell, and each of the approximately 100,000 different kinds of protein in a human cell has a unique structure and function. Elucidating the structures and functions of proteins and structural RNA molecules (for humans and for other organisms) is the central task of molecular biology.

In biosequence databases, there are several pressing data mining tasks, including:

- Find the genes in the DNA sequences of various organisms from among DNA devoted in part to other functions as well. Gene-finding programs, such as GRAIL, GeneID, GeneParser, GenLang, FGENEH, Genie, and EcoParse (see e.g., [6, 7, 9]), use neural nets and other artificial intelligence or statistical methods to locate genes in DNA sequences.[3] Looking for ways to improve the accuracy of these methods is a major thrust of current research in this area.
- Develop methods to search the database for sequences that have higher-order structure or function similar to that of the query sequence, rather than doing a more naive string matching on the sequences themselves. The unique folded structure of each biomolecule (e.g., protein and RNA) is crucial to its function.

Two popular systems for modeling proteins, based on the HMM ideas mentioned earlier, are HMMer and SAM. HMMs and their variants have also been

[3] See "Gene Structure Prediction by Linguistic Methods" by S. Dong and D.B. Searls in *Genomics* (1994) and "Prediction of Gene Structure" by R. Guigo, S. Knudsen, N. Drake, and T. Smith in the *Journal of Molecular Biology* (1994).

applied to the gene-finding problem [6, 7] and to the problem of modeling structural RNA.[4] The gene-finding methods GeneParser, Genie, and EcoParse, mentioned earlier, are examples of this. RNA analysis uses an extension of HMMs called stochastic context-free grammars. This extension permits modeling certain types of interactions among letters of a sequence that are distant in the primary structure but adjacent in the folded RNA structure, a function simple HMMs cannot perform.

Computer-based analysis of biosequences increasingly affects the field of biology. Computational biosequence analysis and database searching tools are now an integrated and essential part of the field, leading to numerous important scientific discoveries in the last few years. Most have resulted from database searches revealing unexpected similarities between molecules previously not known to be related. However, these methods are increasingly important in the direct determination of structure and function of biomolecules as well.

HMMs and related models have been successful in helping scientists with this task because they provide a solid statistical model flexible enough to incorporate important biological knowledge. The key challenge is to build computer methods that can interpret biosequences using a still more complete integration of biological knowledge and statistical methods at the outset, allowing biologists to operate at a higher level in the interpretation process, where their creativity and insight is of maximum value.

Geosciences: Quakefinder and CONQUEST

A major problem facing scientists in such domains as remote sensing is the fact that important signals about temporal processes are often buried within noisy image streams, requiring the application of systematic statistical inference concepts in order for raw image data to be transformed into scientific understanding.

One class of problems that exploit inference in this way is the measurement of subtle changes in images. Consider, for example, the case of two images, taken before and after an earthquake. If the earthquake fault motions are much smaller in magnitude than the pixel resolution (a relatively common scenario), it is essentially impossible to describe and measure the fault motion by simply comparing the two images manually (or even by naive differencing by computer). However, by repeatedly registering different local regions of the two images (a task known to be doable to subpixel precision), it is possible to infer the direction and magnitude of ground motion due to the earthquake. This fundamental concept is broadly applicable to many data mining situations in the geosciences and other fields, including earthquake detection, continuous monitoring of crustal dynamics and natural hazards, target identification in noisy images, and more.

One example of such a geoscientific data mining system is Quakefinder [10], which automatically detects and measures tectonic activity in the Earth's crust by examining satellite data. Quakefinder has been used to automatically map the direction and magnitude of ground displacements due to the 1992 Landers earthquake in Southern California over a spatial region of several hundred square kilometers at a resolution of 10 m to a (sub-pixel) precision of 1 m. It is implemented on a 256-node Cray T3D parallel supercomputer to ensure rapid turnaround of scientific results. The issues of developing scalable algorithms and their implementation on scalable platforms addressed here are in fact quite general and are likely to influence the great majority of future data mining efforts geared to the analysis of genuinely massive datasets.

In addition to automatically measuring known faults, the system permits a form of automatic knowledge discovery by indicating novel unexplained tectonic activity away from the primary Landers faults—activity never before observed. Future work will focus on the measurement of continuous processes over many images, instead of simply abrupt behavior seen during earthquakes, and to related image-understanding problems.

Analysis of atmospheric data is another classic area in which processing and data collection power has far outstripped our ability to interpret the results. The mismatch is huge between pixel-level data and scientific language that understands such spatiotemporal patterns as cyclones and tornadoes. Cross-disciplinary collaborations attempt to bridge this gap, as exemplified by the team formed by JPL and UCLA to develop COncurrent QUErying Space and Time (CONQUEST) [11].

Parallel supercomputers were used in CONQUEST to implement queries concerning the pres-

[4] See "Stochastic Context-Free Grammars for RNA modeling" by Y. Sakakibara, M. Brown, R. Hughey, I.S. Mian, K. Sjolander, R.C. Underwood, and D. Haussler in *Nucleic Acids Research* (1994).

ence, duration, and strength of extratropical cyclones and distinctive blocking features in the atmosphere, scanning through this dataset in minutes. Upon extraction, the features are stored in a relational database. This content-based indexing dramatically reduces the time required to search the raw datasets of atmospheric variables when further queries are formulated. The system also features parallel implementations of singular value decomposition and neural network pattern recognition algorithms in order to identify spatiotemporal features as a whole. The long-term hope is that a common set of flexible, extensible, and seamless tools can be applied across a number of scientific domains.

Conclusions and Challenges

Several issues need to be considered when contemplating a KDD application in science datasets. Some are common with many other data mining applications (e.g., feature extraction, choice of data mining tasks and methods, and understandability of derived models and patterns) [3]. Some considerations are more important in science applications than in financial or business KDD applications, including:

- Ability to use prior knowledge during mining (more documented knowledge is typically available in science applications);
- More stringent requirements for accuracy (e.g., better than 90% accuracy was required for SKICAT);
- Issues of scalability of machines and algorithms (e.g., parallel supercomputers used in scientific applications); and
- Ability to deal with minority (low-probability) classes, whose occurrence in the data is rare, as in SKICAT clustering.

In conclusion, we point out that KDD applications in science may generally be easier than applications in business, finance, or other areas—mainly because science users typically know their data in intimate detail. This knowledge allows them to intuitively guess the important transformations. Scientists are trained to formalize intuitions into procedures and equations, making migration to computers easier. Background knowledge is usually available in well-documented form (papers and books), providing backup resources when the initial data mining attempts fail. This luxury (sometimes a burden) is not usually available in nonscientific fields. ◘

References

1. Burl, M.C., Fayyad, U., Perona, P., Smyth, P., and Burl, M.P. Automating the hunt for volcanoes on Venus. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR-94)* (Seattle 1994). IEEE Computer Science Press, Los Alamitos, Calif., 1994, pp. 302–308.
2. Chothia, C. One thousand families for the molecular biologist. *Nature 357 (*1992), 543–544.
3. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. *Advances in Knowledge Discovery in Databases.* MIT Press, Cambridge, Mass., 1996.
4. Kennefick, J.D., DeCarvalho, R.R., Djorgovski, S.G., Wilber, M.M., Dickinson, E.S., Weir, N., Fayyad, U., and Roden, J. *Astron. J. 110,* 1 (1995), 78–86.
5. Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol. 235* (1994), 1501–1531.
6. Krogh, A., Mian, I.S., and Haussler, D. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res. 22* (1994), 4768–4778.
7. Kulp, D., Haussler, D., Reese, M., and Eeckman, F. A generalized hidden Markov model for the recognition of human genes in DNA. In *Proceedings of the Conference on Intelligent Systems in Molecular Biology* (1996). AAAI Press, Menlo Park, Calif., 1996.
8. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE 77 (*1989), 257–286.
9. Snyder, E.E., and Stormo, G.D. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucleic Acids Res. 21 (*1993), 607–613.
10. Stolorz, P., and Dean, C. Quakefinder: A scalable data mining system for detecting earthquakes from space. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (Portland, Oreg., 1996), AAAI Press, Menlo Park, Calif., 1996.
11. Stolorz, P., Nakamura, H. Mesrobian, E., Muntz, R.R., Shek, E.C., Mechoso, C.R., Farrara, J.D. Fast spatiotemporal data mining of large. geophysical datasets. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining* (Montréal, Aug. 1995), AAAI Press, Menlo Park, Calif. 1995, pp. 300–305.
12. Weir, N., Fayyad, U.M., and Djorgovski, S.G. Automated star/galaxy classification for digitized POSS-II. *Astron. J. 109,* 6 (1995), 2401–2412.
13. Weir, N., Djorgovski, S.G., and Fayyad, U.M. Initial galaxy counts from digitized POSS-II. *Astron. J. 110,* 1 (1995), 1–20.

Additional references for this article can be found at http://www.research.microsoft.com/research/datamine/CACM-DM-refs/.

USAMA FAYYAD is senior researcher at Microsoft and a Distinguished Visiting Scientist at the Jet Propulsion Laboratory, California Institute of Technology. He can be reached at fayyad@microsoft.com.

DAVID HAUSSLER is a professor of computer science at the University of California, Santa Cruz. He can be reached at haussler@cse.ucsc.edu.

PAUL STOLORZ is technical group supervisor at the Jet Propulsion Laboratory, California Institute of Technology. He can be reached at pauls@aig.jpl.nasa.gov.