Data Mining and Knowledge Discovery in Databases: Applications in Astronomy and Planetary Science

Usama M. Fayyad*

Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
fayyad@microsoft.com
http://www.research.microsoft.com/research/dtg

* Author is also affiliated with:
Machine Learning Systems Group
Jet Propulsion Laboratory
California Institute of Technology
http://www-aig.jpl.nasa.gov/mls

Abstract of Invited Talk

Overview of the Topic

Knowledge Discovery in Databases (KDD) is a new field of research concerned with the extraction of high-level information (knowledge) from low-level data (usually stored in large databases) [1]. It is an area of interest to researchers and practitioners from many fields including: AI, statistics, pattern recognition, databases, visualization, and high-performance and parallel computing. The basic problem is to search databases for patterns or models that can be useful in accomplishing one or more goals. Examples of such goals include:

- prediction (e.g. regression and classification),
- descriptive or generative modeling (e.g. clustering),
- data summarization (e.g. report generation), or
- visualization of either data or extracted knowledge (e.g. to support decision making or exploratory data analysis).

KDD is a process that includes many steps. Among these steps are: data preparation and cleaning, data selection and sampling, preprocessing and transformation, data mining to extract patterns and models, interpretation and evaluation of extracted information, and finally evaluation, rendering, or use of final extracted knowledge. Note that under this view, *data mining constitutes one of the steps of the overall KDD process*. The other steps are essential to make the application of data mining possible, and to make the results useful. Within data mining, methods for deriving patterns or extracting models originate from statistics, machine learning, statistical pattern recognition, uncertainty management, and database methods such as on-line analysis processing (OLAP) or association rules [2].

The process is typically highly interactive and may involve many iterations before useful knowledge is extracted from the underlying data. This talk will give an overview and summary of the rapidly growing field of KDD, and then focus on two specific applications in scientific data analysis to illustrate the potential, limitations, challenges, and promise of KDD. An overview of the KDD process is given in [3].

Science Data Analysis

Today's science instruments are capable of gathering huge amounts of data, making traditional human-based comprehensive analysis an infeasible endeavor. This has been a primary motivation to develop tools to automate science data analysis tasks. The talk will describe efforts to develop a new generation of data mining systems where users specify what to search for simply by providing the system with training examples, and letting the system automatically learn what to do. The system would then automatically sift through the data and catalog objects of interest for analysis purposes.

The learn-from-example approach is a natural solution to a problem we call the *query formulation problem* in the exploration and analysis of image data [4]: How does one express a query for objects that are typically only recognized by visual intuition? Translating human visual intuition to pixel-level algorithmic constraints is a difficult problem. By asking the user to simply "show" the system examples of objects of interest, then let the system figure out how to formulate the appropriate query, we believe the problem can be surmounted in certain circumstances.

Two applications at JPL will be used to illustrate the learning techniques and their effects. The first targets automating the cataloging of sky objects in a digitized sky survey consisting of three terabytes of image data and

containing on the order of two billion sky objects. The Sky Image Cataloging and Analysis Tool (SKICAT) [5] allows for automated and accurate classification, enabling the automated cataloging of an estimated two billion sky objects, the majority of which being too faint for visual recognition by astronomers. This represents an instance where learning algorithms solved a significant and difficult scientific analysis problem. Several new results in astronomy have been achieved based on the SKICAT catalog [6]. Recent results of the application of SKICAT to help in discovery of new objects in the Universe include the discovery of 16 new high-redshift quasars: some of the furthest and oldest objects detectable by today's instruments [7].

The second system we describe is called JARtool (JPL Adaptive Recognition Tool) [8]. JARtool is being initially developed to detect and catalog an estimated one million small volcanoes (< 15km in diameter) visible in a database consisting of over 30,000 images of the planet Venus. The images were collected by the Magellan spacecraft using synthetic aperture radar (SAR) to penetrate the permanent gaseous cloud cover that obscures the planet's surface in the optical range.

Work at JPL's Machine Learning Systems Group continues to extend data mining techniques to automate analysis in other areas of science including: cataloging of Sun spots, remote-sensing detection of earthquake faults [9], spatio-temporal analysis of atmospheric data, and others (see http://www-aig.jpl.nasa.gov/mls/ for live descriptions of ongoing work).

Other Applications

Although this talks focuses on applications in science data analysis, we believe these techniques to be applicable to a wide range of problems and have little to do with the fact that the data happens to be images. Potential applications include medical imaging, automated inspection and diagnosis in manufacturing, decision support systems, database marketing, and summarization/visualization of large databases. Coverage of applications in science data analysis is given in [10]. Coverage of industrial applications of KDD is provided in [11].

Beware the Hype

While it is true that in many situations some simple data mining work can result in great successes, this by no means justifies the public perception that data mining can be used to solve all analysis problems. The fundamental problems in the field are far from being solved. The basic problem in KDD is one of statistical inference from a finite set of data. Issues of model overfitting and justification of findings remain as major challenges. Many models are extractable from data (in fact infinitely many models from a finite data set). Most of these are likely to be due to spurious correlations and random chance (simply because the data is finite, and computation is limited). See [12] for a critique and warnings of pitfalls. Hence, any derived predictions and any inference of causality information from data is to be taken with a fairly huge helping of salt. Furthermore, effective techniques for incorporating the necessary prior knowledge about an application into a data mining algorithm to help the automated system avoid some of the basic traps are still lacking.

The fact that initial successes exist should be taken as encouraging signs that this new and emerging field holds some promise to address the daunting problems of information overload facing modern society.

Information Resources

A good starting point to get a summary of what has been done in this field is to start with [1] and [13]. Appendix 1 of [1] by Kloesgen and Zytkow provides a glossary of terminology used in KDD, while Appendix 2 by Piatetsky-Shapiro provides pointers to various resources. The next source of information is to follow up on the following resources on the internet and the world-wide web:

- http://www-aig.jpl.nasa.gov/kdd95/ is the homepage of the First International Conference on Knowledge Discovery and Data Mining (KDD-95); The homepage of KDD-96 is at http://www-aig.jpl.nasa.gov/kdd96. Information on the new journal for this field: *Data Mining and Knowledge Discovery* can be obtained at http://www.research.microsoft.com/research/datamine.
- http://info.gte.com/~kdd/ is the *Knowledge Discovery Mine* maintained by Gregory Piatetsky-Shapiro at GTE Laboratories. This site serves as a collecting point of information directly relevant to KDD, including tools available free on-line, commercial products, pointers to many other relevant groups and organizations, as well as a repository of all back issues of *KDD Nuggets.*
- *KDD Nuggets:* is a moderated mailing list that serves as the main forum of communicating with or receiving news of interest to the KDD community (kdd-request@gte.com).

Acknowledgments

The applications described in this talk were conducted while the author was still with the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Collaborators on the SKICAT work at Caltech and Palomar Observatory include S.G. Djorgovski and Nick Weir. We also thank Reinaldo DeCarvalho and Julia Kennefick (Caltech), Joe Roden, John Loch, Maureen Burl, Scott Burleigh, Jennifer Yu, and Alex Gray (JPL).

JARtool work is in collaboration with Padhraic Smyth (JPL), Michael Burl and Pietro Perona (Caltech), and Jayne Aubele and Larry Crumpler (Brown University). We also thank Maureen Burl, Joe Roden, Victoria Gor, and Michael Turmon (JPL).
Program management of this work at NASA was under the direction of Melvin Montemerlo (Code XS), and at JPL under direction of David Atkinson and Richard Doyle.
In my overview coverage of data mining and KDD, I have borrowed material from my co-authors on [3]: Gregory Piatetsky-Shapiro and Padhraic Smyth.

References

[1] Fayyad U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. (Eds.) 1996. *Advances in Knowledge Discovery and Data Mining.* Cambridge, Mass.: MIT Press/AAAI Press.

[2] Agrawal, A.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, I. 1996. "Fast Discovery of Association Rules", in *Advances in Knowledge Discovery and Data Mining.* Fayyad U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. (Eds.), Cambridge, Mass.: MIT Press/AAAI Press.

[3] Fayyad U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. "From Data Mining to Knowledge Discovery: An Overview", in *Advances in Knowledge Discovery and Data Mining.* Fayyad U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. (Eds.), Cambridge, Mass.: MIT Press/AAAI Press.

[4] Fayyad, U. and Smyth P. 1995. "The Automated Analysis, Cataloging, and Searching of Digital Libraries: A Machine Learning Approach", in *Digital Libraries, Lecture Notes in Computer Science 916,* N.R. Adam, B.K. Bhargava, and Y. Yesha (Eds.) Berlin: Springer-Verlag.

[5] Fayyad U.; Djorgovski, S.G.; and Weir, N. 1996. "Automating the Analysis and Cataloging of Sky Surveys", in *Advances in Knowledge Discovery and Data Mining.* Fayyad U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. (Eds.), Cambridge, Mass.: MIT Press/AAAI Press.

[6] Weir, N.; Djorgovski, S.G.; and Fayyad, U.M. 1995. "Initial Galaxy Counts From Digitized POSS-II", *Astronomical Journal,* 110-1:1-20.

[7] Kennefick, J.D.; De Carvalho, R.R.; Djorgovski, S.G.; Wilber, M.M.; Dickinson, E.S.; Weir, N.; Fayyad, U.; and Roden, J. (1995). "The Discovery of Five Quasars at z>4 using the Second Palomar Sky Survey", *Astronomical Journal,* 110-1:78-86.

[8] Burl, M.C.; Fayyad, U.; Perona, P.; Smyth, P.; and Burl, M.P. (1994). "Automating the Hunt for Volcanoes on Venus", in *Proc. of Computer Vision and Pattern Recognition Conference (CVPR-94),* pp. 302-308, IEEE Computer Society Press.

[9] Stolorz, P. and Dean, C. 1996. "Quakefinder: A Scalable Datamining System for Detecting Earthquakes from Space", submitted to *Second International Conf. on Knowledge Discovery and Data Mining.* AAAI Press.

[10] Fayyad, U.; Haussler, D.; and Stolorz, P. 1996. "Science Applications", special issue on Data Mining and Knowledge Discovery, *Communications of the ACM,* forthcoming.

[11] Brachman. R.J.; Khabaza, T.; Kloesgen, W.; Piatetsky-Shapiro, G.; and Simoudis, E. 1996. "Industrial Applications of Data Mining and Knowledge Discovery", in *Proc. Of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96),* Menlo Park, Calif.: AAAI Press.

[12] Glymour, C.; Madigan, D.; Pregibon, D.; and Smyth, P. 1996. "Statistics and Data Mining", special issue on Data Mining and Knowledge Discovery, *Communications of the ACM,* forthcoming.

[13] Piatetsky-Shapiro, G. and Frawley W. (1991). *Knowledge Discovery in Databases.* Cambridge, MA: MIT Press.