2000

Data Mining: A Brief Introduction to the Field and Research Community

William E. Spangler
Duquesne University, wspangle@wvu.edu

H. Michael Chung
California State University - Long Beach, hmchung@csulb.edu

Fredric C. Gey
University of California - Berkeley, gey@ucdata.berkeley.edu

Data Mining: A Brief Introduction to the Field and Research Community

William E. Spangler, Duquesne University, wspangle@wvu.edu
H. Michael Chung, California State University, Long Beach, hmchung@csulb.edu
Fredric C. Gey, University of California, Berkeley, gey@ucdata.berkeley.edu

Abstract

As organizations seek to understand and exploit vast – and increasing -- amounts of information brought by new technologies and practices, research in data mining and knowledge discovery is becoming increasing important. This paper presents a brief overview of the field of data mining, including research issues and information resources available through the research community.

Overview

With the advent of new information technologies and business practices, including the explosion of electronic commerce and the trend toward mass personalization, businesses are accumulating massive, and increasing, amounts of data. Surveys indicate that, in recent years, many organizational databases have doubled, tripled or even quadrupled each year – a trend that is expected to continue and perhaps accelerate (Winter, 1999). Because the data contained in those databases potentially is of enormous value, businesses are motivated to search for ways to make sense of it. Data mining allows organizations to make sense of data by uncovering useful patterns that indicate, for example, consumer buying habits, competitor strategies, credit worthiness, incidence of fraud, computer network intrusion, and other information of strategic importance. As a result, the field is becoming increasingly important to both researchers and practitioners.

Data mining is closely related to the field of machine learning, and generally is considered to be an umbrella term describing work in a number of related fields and methodologies -- including statistics, AI, neural networks, visualization, tree/rule induction, database techniques, and many others. Within the context of data mining, each of these methodologies is used to infer useful patterns in collected data, where a pattern describes the relationship among variables as a set of conditional probabilities. (For example, given that a customer buys expensive cigars, s/he also is likely also to buy expensive cars.)

Fayyad et al suggest that the two primary goals of data mining in practice are *prediction* and *description* (Fayyad, et al., 1996). Prediction entails using a set of variables in a data set to infer some unknown value or values, while description entails finding 'human-interpretable' patterns that might characterize the data. These general goals in turn can be achieved through various data mining approaches. Weiss and Indurkyha suggest three categories of approaches: 1. *math-based* approaches, which include linear discriminant analysis and neural networks, 2. *distance-based* approaches such as k-nearest neighbor, and 3. *logic-based* approaches, which include decision tree or rule induction (Weiss and Indurkhya, 1998).

Data mining draws heavily from the field of empirical or inductive learning, which seeks to induce rules or procedures from a set of examples (Shavlik and Dietterich, 1990). Inductive learning can be further divided into unsupervised and supervised learning. In unsupervised learning, a learning algorithm is not informed about the classes of individual input cases. Instead, the algorithm attempts to group -- and therefore classify -- the cases based on a similarity metric derived from case attributes. Unsupervised learning is composed of various cluster analysis and knowledge discovery methods. In supervised learning, by contrast, a learning algorithm is provided examples of both the attributes of each case and the class to which each case has been assigned. From this information the algorithm is expected to construct a generic function that can be used to classify and predict the outcomes of new examples. Techniques such as neural networks, decision tree induction and linear discriminant analysis employ supervised learning (Chung and Silver, 1992). Most regression techniques, as well as Bayesian classification techniques, similarly employ supervised learning approaches.

Bases for comparing data mining techniques

Considering the number of criteria by which one can measure the performance of a data mining technique, it is not surprising that no particular technique is universally superior to the others. Because research has shown that performance of data mining techniques tends to be dependent on the situation and on the goals of the end user, judging the relative performance of data mining methods requires consideration of each of the various measures of a method's performance or worth (Kim, et al.,1998). They are:

Predictive accuracy. Many comparative studies of data mining methods have measured the predictive accuracy and error rate of each method (e.g., see Chung and Tam, 1993; Spangler, et al., 1999). However, even on this relatively narrow basis, there has been little consensus regarding performance across methods. Performance is highly dependent on the domain and setting, the size and nature of the data set, the presence of noise and outliers in the data, and the validation technique(s) used.

Comprehensibility. Henery uses this term to indicate the need for a classification method to provide clearly understood and justifiable decision support to a human manager or operator (Henery, 1994). In this regard, decision tree and rule induction systems, because they explicitly structure the reasoning underlying the classification process, tend to have an inherent advantage over both traditional statistical classification models and neural networks. Tessmer et al argue that while the traditional statistical methods provide efficient predictive accuracy, they do not provide an explicit description of the classification process (Tessmer, et al., 1993). Weiss and Kulikowski (Weiss and Kulikowski, 1991) suggest that any explanation resident in mathematical inferencing techniques is buried in computations which are inaccessible to the 'mathematically uninclined'. As a result, these techniques hold the potential to engender misunderstandings and therefore misuse of the results. Rules and decision trees, on the other hand, are more compatible with human reasoning and explanations.

Speed of training and classification. Speed can be an important consideration in some situations. Henery suggests that a number of real-time applications, for example, must sacrifice some accuracy in order to classify and process items in a timely fashion (Henery, 1994). Again, because of situational dependencies, it is difficult to make generalizations about the computational expense of each method -- although back-propagation in neural networks, in particular, is a notoriously time-consumptive process (Weiss and Kulikowski, 1991).

Selection of attributes. The attributes selected for consideration and their relative importance in influencing the outcome also is an indication of the performance of a method. This concept of *diagnostic validity* of induction methods was proposed by Currim et al (Currim, et al., 1986), and was used by Messier and Hansen to compare the attributes selected by each of their induction methods (Messier and Hansen, 1988).

Research issues & directions

The field of data mining is rich in research and application development opportunities. According to John, many of the opportunities come from two broad areas: 1) enhancing current data mining approaches through cutting edge research from statistics, machine learning, visualization and database management, and 2) making data mining products and tools easier for domain experts to use (John, 1997). Specific opportunities from the first area include the development of new and hybrid algorithms, and continuing research into mining of new data formats and structures. Newer formats would include multimedia data such as image, video and sound.

Equally important is the issue of usability, which Fayyad argues is the essential goal of data mining research and development (Fayyad, 1999). He notes that the construction of better and more accurate algorithms is not beneficial unless end users are able to configure and run the algorithms conveniently, understand the results, and use the results for the solution of relevant problems. Consequently, issues pertaining to development and use of data mining tools, and management of the data mining process, become especially important. For example, the continuing integration of data mining tools and database systems is important because, as Fayyad notes, integration simplifies the data mining tasks of *model deployment* and *data maintenance and management* -- which are two of the costliest components of the data mining process.

Other costly aspects of data mining include the 'pre-processing' tasks of selecting, extracting, cleaning and preparing the data prior to data analysis. End user support in this area can come in part from research exploring the automation of *data cleaning*. John suggests that although current algorithms are incapable of discovering and making independent decisions about problem data, more modest algorithms that can identify and flag 'suspicious' data for the user can be developed. (John, 1997) Research also can support users in the areas of data and attribute selection, in part through the incorporation of knowledge-based planning techniques that draw upon specific domain knowledge (Waltz and Hong, 1999), and in choosing appropriate algorithms given a particular problem and situation description. A knowledge-based model might also support a user in the interpretation of data mining results post hoc.

Other salient research issues do not necessarily fall into one of John's two categories listed above. For example, performance evaluation of data mining approaches is an issue related to algorithm enhancement as well as to usability. As mentioned in Chung and Gray (1999), comparative research must address the various criteria for evaluation, which are somewhat complex, unstructured, and problem-specific. From a business and managerial perspective, data mining research also can explore data mining and organizational goal alignment; i.e., what goals does data mining address, and how do those goals contribute to overall IT and organizational goals? Business issues also include the impact of the internet and electronic commerce, which have resulted in the accumulation of vast amounts of transactional and demographic data. The motivation to exploit this data has spurred an increased interest in e-commerce-based data warehouses, and the introduction of a new term for data mining in this area -- i.e., 'web mining'. Piatetsky-

Shapiro notes that the collection of vast amounts of data over relatively short periods of time – which results in relatively little historical information being available -- requires the development of new strategies for making predictions from data (Piatetsky-Shapiro, 1999). An example of these new strategies is *collaborative filtering*, which predicts a person's future behavior by drawing on information about the historical behavior of *other* people.

Research community & resources

An excellent general source of information about data mining research, applications and pedagogy is presented at the 'KDNuggets' web site (www.kdnuggets.com). KDNuggets, which spawned from Gregory Piatetsky-Shapiro's efforts at GTE in the mid-1990s, contains information and links to data mining software, publications, courses, data sets and other data mining-related web sites. The KDNuggets site also provides access to *KDNuggets News*, a free twice-monthly e-newsletter with over 8000 subscribers. Other web-based sources of information include David Aha's exceptionally extensive list of Machine Learning Resources (www.aic.nrl.navy.mil/~aha/research/machine-learning.html), the Machine Learning Network (MLNet) Online Information Service (www.mlnet.org/), the European Cross Industry Standard Process for Data Mining (CRISP-DM) project (http://www.crisp-dm.org/), and the Online Machine Learning Resources list maintained by the ML Group at the Austrian Research Institute for Artificial Intelligence (http://www.ai.univie.ac.at/oefai/ml/ml-resources.html).

A number of professional organizations have been established to provide a forum for data mining research and practice. One prominent example is the Association for Computing Machinery (ACM) Special Interest Group on Knowledge Discovery and Data Mining - SIGKDD (www.acm.org/sigkdd). As described in the SIGKDD home page, the group seeks to encourage the basic research in KDD (through annual research conferences, newsletter and other related activities), adoption of 'standards' in the market in terms of terminology, evaluation, methodology, and interdisciplinary education among KDD researchers, practitioners, and users. Like KDNuggets, SIGKDD also publishes an electronic newsletter, called *SIGKDD Explorations*. Other organizations with an interest in data mining include: the American Association for Artificial Intelligence (www.aaai.org), the ACM Special Interest Group on Management of Data – SIGMOD (www.acm.org/sigmod/), the ACM Special Interest Group on Artificial Intelligence – SIGART (sigart.acm.org), the Technical Committee on Data Engineering (TCDE) of the IEEE Computer Society (www.ccs.neu.edu/groups/IEEE/tcde/index.html), the INFORMS Computing Society (http://www.informs.org/subdiv/Society/ICS/), the INFORMS College on Artificial Intelligence (www.informs.org/subdiv/Section/ai.html), and the Society for Artificial Intelligence and Statistics (www.vuse.vanderbilt.edu/~dfisher/ai-stats/society.html). In addition to these professional organizations, a number of private academic and industrial organizations have established research and development centers in data mining and machine learning. Each of the general DM information sources mentioned above provides a list of academic research centers around the world.

A number of IS and AI-related journals and conferences focus on data mining and knowledge discovery. The primary journal devoted exclusively to the topic is *Data Mining and Knowledge Discovery*, which started publication in 1997. Other journals concentrating on data mining include *Intelligent Data Analysis*, *Machine Learning*, and *IEEE Transactions on Knowledge and Data Engineering*. A number of other IS/IT journals have published special issues on data mining, including the Journal of Management Information Systems (Vol. 16, No. 1, 1999), IEEE Intelligent Systems (Vol. 14, No. 6, 1999), Communications of the ACM (Vol. 42, No. 11 November, 1999) and IEEE Computer (Vol. 32, No. 8, 1999). The primary conference for the data mining research community is the ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*, which began in 1995 under the sponsorship of AAAI and other organizations. In addition, most mainstream IS conferences – including HICSS, ICIS, AMCIS, INFORMS and DSI -- have tracks or sessions on data mining or related areas. HICSS has kept Data Mining and Knowledge Discovery track since 1998, and is expanding to the Information Retrieval area.

Conclusions

A number of trends suggest that the problems of data collection, organization and analysis will not be solved anytime soon. Advances in information technology – specifically networking and storage technologies – will allow faster accumulation of ever-increasing amounts of data. Electronic commerce, aided by new hardware and software technologies, and new business models, will accelerate information interchange between and among businesses and consumers. Continuing advances in scientific research, such as the Human Genome Project and other biotechnology endeavors, likewise will produce vast amounts of data of various types and structures – all of which must be analyzed and explored. Considering the opportunities that these trends present to researchers, the future of data mining clearly is very bright.

References

Chung, H.M. and Silver, M. "Rule-based Expert Systems and Linear Models: An Empirical Comparison of

Learning-By-Examples Methods," *Decision Sciences*, Vol.23, No.3, 1992, pp. 687-707.

Chung, H.M. and Tam, K.Y. "A Comparative Analysis of Inductive Learning Algorithms," *Intelligent Systems in Accounting, Finance and Management* (2:1), 1993, pp. 3-18.

Chung, H.M. and Gray, P. "Critical Issues in Data Mining," *Journal of Management Information Systems (16:1)*, 1999, pp 11-16.

Currim, I.S., Meyer, R.J. and Le, N. "A Concept-Learning System for the Inference of Production Models of Consumer Choice," Working Paper UCLA, 1986.

Fayyad, U. "Editorial," *SIGKDD Explorations* (1:1), www.acm.org/sigkdd, 1999.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. "From Data Mining to Knowledge Discovery in Databases," In *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Ed.), AAAI Press, Menlo Park, CA, 1996.

Henery, R.J. "Classification," In *Machine Learning, Neural and Statistical Classification*, D. Michie, D. J. Spiegelhalter and C. C. Taylor (Ed.), Ellis Horwood, New York, 1994, pp. 6-16.

John, G.H. "Enchancements to the Data Mining Process," Stanford University, 1997.

Kim, C., Chung, H.M., and Paradice, D. "Inductive Modeling of Expert Decision-Making in Loan Evaluation: A Decision Strategy Perspective," *Decision Support Systems*, Vol. 21, No. 2, 1998, pp. 83-98.

Messier, W.F. and Hansen, J.V. "Inducing Rules for Expert System Development: An Example Using Default and Bankruptcy Rules," *Management Science* (34:12), 1988, pp. 1403-1415.

Piatetsky-Shapiro, G. "The Data Mining Industry Coming of Age," *IEEE Intelligent Systems* (14:6), 1999, pp. 32-34.

Shavlik, J.W. and Dietterich, T.G. "General Aspects of Machine Learning," In *Readings in Machine Learning*, J. W. Shavlik and T. G. Dietterich (Ed.), Morgan Kaufmann, San Mateo, CA, 1990, pp. 1-10.

Spangler, W.E., May, J.H. and Vargas, L.G. "Choosing Data Mining Methods for Multiple Classification: Representational and Performance Measurement Implications for Decision Support," *Journal of Management Information Systems* (16:1), 1999.

Tessmer, A.C., Shaw, M.J. and Gentry, J.A. "Inductive Learning for International Financial Analysis: A Layered Approach," *Journal of Management Information Systems* (9:4), 1993, pp. 17-36.

Waltz, D. and Hong, S.J. "Data Mining: A Long-Term Dream," *IEEE Intelligent Systems* (14:6), 1999, pp. 30-31.

Weiss, S.M. and Indurkhya, N. *Predictive Data Mining*, Morgan Kaufmann, San Francisco, CA, 1998.

Weiss, S.M. and Kulikowski, C.A. *Computer Systems That Learn*, Morgan Kaufmann, San Francisco, 1991.

Winter, R. "Saving More," *Intelligent Enterprise*, (2:6), 1999.