KDD-93: Progress and Challenges in Knowledge Discovery in Databases

Gregory Piatetsky-Shapiro (GTE Laboratories)*, Christopher Matheus (GTE Laboratories), Padhraic Smyth (J]'].), Ramasamy Uthurusamy (GM Research I Laboratories)

November 16, 1993

Interest in Knowledge Discovery in Databases (KDD) continues to increase, driven by the rapid growth in the number and size of large databases and the applications-driven demand to make sense of them. The research side of KDD is of growing interest to researchers in machine learning, statistics, intelligent databases, and knowledge acquisition, as evidenced by the number of recent workshops (l'iatetsky-Shapiro 1991a, 1991b, Zytkow 1992, Ziarko 1993) and special journal issues (I'iatetsky-Shapiro 1992, Zytkow 1993, Cercone 1993) devoted to or closely related to discovery in databases. The application side is of interest to any business or organization with large databases. KDD applications have been reported in many areas of business, government, and science (Piatetsky-Shapiro and Frawley 1991, Innon and Osterfelt 1991, Parsaye and Chignell 1993).

The notion of discovery in databases has been given various names, including knowledge extraction, data mining, database exploration, data pattern processing, data archaeology, information harvesting, siftware, and even (when done poorly) data dredging. Whatever the name, the essence of KDD is tile *nontrivial extraction of* implicit, previously *unknown, and potentially useful information from data* (Frawley et al 1992). KDD encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies (see Matheus et al 1993).

Over 60 researchers from 10 countries took part in the third KDD workshop (Piatetsky-Shapiro 1993) held during AAAI-93 in Washington, D.C, in the simmering July heat. A major trend evident at the workshop was the transition to applications in the core KDD area of discovery of relatively simple patterns in relational databases; the most successful applications are appearing in the areas of greatest need, where the databases are so large that the manual analysis is impossible. Progress was also facilitated by the availability of commercial KDD tools, both for generic discovery and for domain-specific applications such as marketing. . At the same time, progress is slowed by problems such as lack of statistical rigor, overabundance of patterns, and poor integration.

Besides applications, the main themes of this workshop were the Discovery of Dependencies and Models and Integrated and interactive KDD Systems.

Real-World A pplications

The applications presented at the workshop fell into three broad application) areas: scientific, financial, and manufacturing. Most of the systems performed some form of classification, while two systems dealt with detecting and describing changes. In addition to talks and poster presentations, several demonstrations of research and commercial discovery systems were given at the workshop.

Scientific applications: Two applications were presented in the area of astronomy. Usama Fayyad (J 1'1,), started the workshop with a talk on Sky image Cataloging and Analysis Tool (SKICAT), an automated

*The authors can be reached at gps@gte.cent, matheus@gte.com, pjs@galway.Jpl.Nasa.Gov, aad samy@gmr.com, respectively

system for analyzing large-scale sky surveys. The multi-terabyte size of the database ruled out a manual approach to image classification. Using a number of innovative machine learning methods, Usama and his colleagues were able to recognize objects at least one magnitude fainter in resolution than was previously possible while achieving an accuracy of about 94%. This work is noteworthy as a real application of machine learning to a difficult problem with results that arc being used try scientists on a daily basis. Padhraic Smyth, also from JPL, gave a related talk on the problem of Image Database Exploration, describing collaborative work with Usama Fayyad. Padhraic described challenging issues in image analysis such as how to measure the right attributes, the role of prior knowledge, incremental learning, and the use of multi-sensor data. Ile also examined how these issues are handled in current JPL tasks such as the analysis of Venus images obtained by the Magellan spacecraft.

Financial applications: Two systems were presented for detecting and describing changes in large business databases. Tej Anand described A .C. Nielsen's recent work on a commercial product called Opportunity Explorer. This system is a redesign and extension of their Spotlight (Anand and Kahn 1992) product for identifying and reporting on trends and exceptional events in the extremely large supermarket sales databases. An innovative feature of Spotlight is the automatic explanation of relationships between key events. Opportunity Explorer is a more general tool for developing interactive, hypertextual reports using knowledge discovery templates which convert a large data space into concise, irltcr-linked information frames. It is marketed to help sales analysts and product managers of consumer packaged goods companies develop better sales strategies.

Christopher Matheus and Gregory Piatetsky-Shapiro of GTE Laboratories presented their Key Findings Reporter (KEFIR), a system for discovering and explaining "key findings" in large relational databases. While the system's design is domain independent, the current focus is on trend and normative analysis of health-care information. KEFIR performs an automatic drill-down on the data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to norms or previous values. It then identifies explanatory relationships between findings, and generates a report using natural language templates and graphics. A prototype of KEFIR has been implemented in C and tcl with an embedded SQL interface.

Three other financial applications used classification methods in the areas of insurance, marketing, and stock market analysis. John Major (Travelers) analyzed the important problem of selecting the most interesting rules among those discovered in data. Ile presented a rule refinement strategy which defined rule "interestingness" via rule accuracy, coverage, simplicity, novelty, and significance. His method gave preference to rules not dominated in these measures by other rules, and removed those that were potentially redundant. in an application of the method to a tropical storm database, the system reduced 161 rules generated try IX], (a product of IntelligenceWare, Inc) to the 10 most interesting ones which were meaningful to a meteorologist.

Wojtek Ziarko (U. of Regina, Canada) presented an application of Reduct Systems' Datalogic/R discovery tool to identify strong predictive rules in stock market data. Month I y data collect cd over a ten year period was analyzed to identify dominant relationships among fluctuations of market indicators and stock prices. Evaluation, by a domain expert, of the results (including both precise and imprecise, strong and weak rules) revealed that the strong rules confirm expert's experiences while weak rules were difficult to interpret. Datalogic/R, a commercially available tool which derives rules using the variable precision rough sets approach, was also demonstrated at the workshop.

Pierro Bonissone and Lisa Rau of GE R&D presented preliminary results of applying decision trees and logistic regression to a database of accounting, customer, and sales information. Initial results that suggest emerging markets and provide feedback on sales performance are encouraging enough to warrant further pursuit of this work.

Manufacturing: Two applications dealt with semiconductor manufacturing and software engineering. Sharad Saxena of Texas Instruments presented his approach to fault isolation during semiconductor manufacturing using automated discovery from wafer tracking databases. These databases contain the history of the semiconductor wafers as they undergo various processing steps. A gcncratc-aad-test approach is taken

for using such databases for automated diagnosis. Based on prior manual analysis of such databases, classes of queries to the database as well as patterns in the responses to these queries that are useful for fault isolation are identified. Diagnosis is accomplished by automating the query generation and the detection of potentially useful patterns. A prototype system was implemented and tested on real data, finding both known and previously unknown faults.

Inderpal Bhandari of IBM presented Attribute Focusing, a method for exploratory analysis of attribute-valued data intended for use by domain experts who do not have a background in data analysis. The approach uses a model of interest, ingness based on magnitude of data values, association of data values, and basic knowledge of the limits of human information processing capabilities, as well as a model of interpretation to guide the domain specialist to discover knowledge from attribute-valued data. This approach has been used successfully by software managers, developers, and testers at IBM to make real-time improvements on their products, as well as on their process of production. Attribute Focusing approach is being used in several IBM laboratories, with reported net savings of hundreds of person days. A I'C-based implementation of the Attribute Focusing approach was demonstrated by Bhandari and Michael Herman of IBM.

Discovery of Dependencies and Models

The second major theme of the workshop was discovery of dependencies and models. The workshop provided clear evidence of the diversity of technical approaches which arc being applied to the general KDD problem. The focus was on the use of particular mathematical and statistical methods for the induction of qualitative relationships directly from data.

Jan Zytkow (Wichita State U.) outlined the latest developments in his joint research with Robert Zembowicz on deriving equations from data. He proposed a computationally simple test for the absence of functional dependency which can eliminate the much more expensive search to determine the form of dependency. The test relics on search for discretization of data into optimal intervals. Initial experiments with their 49er system showed that the test significantly reduces the computation time, while losing only a few actual equations, typically those with a particularly poor fit to data.

Dependency networks are an important form of discovered knowledge, and recent progress in this field (Pearl 1992, Spirtes et al 1993) is very encouraging for KDD. Probabilistic networks area powerful knowledge representation medium, providing a bridge between the power of explicit knowledge representation in graphical form and more subtle (but robust) quantitative statistical methods. Greg Cooper (U. of Pittsburgh), presented the latest results in his research on the usc of Bayesian statistical methods for the learning of causal probabilistic network models that contain hidden variables. In earlier work, Cooper has demonstrated that networks with hidden variables can be directly inferred from data. in this talk, he showed how to structure the calculations to dramatically speed up the computation.

Cooper also summarized recent research progress relevant to the discovery of directed probabilistic networks from data: there is a greater understanding of what relationships can be captured from data by directed acyclic graphs (DAGs) and which DAGs arc indistinguishable based only on data; new methods have been developed for the discovery of probabilistic networks with measured and possibly unmeasured (latent) variables; these methods have been applied to real data with promising results. The major improvements needed for applications to real databases are computational (search) efficiency, integration of different methods, especially those dealing with discrete and continuous variables, and estimating the confidence and the stability of the output.

Sašo Džeroski (Jožef Stefan Institute, Ljubljana, Slovenia) gave an invited overview of Inductive Logic Programming (ILP) methods for KDD. ILP is an important paradigm that goes beyond the typical attribute-value relations (which are the limit of what can be learned by most current machine learning methods) to the more genera] language of flrstl-order relations. The field has developed rapidly in recent years (Muggleton 1 992), and now boasts relatively sophisticated algorithms and methods for handling a variety of problems, with great potential for KDD applications (Lavrač and Džeroski 1 993). Džeroski out lined the motivation for ILP and proceeded in his talk from early work through more recent extensions and up to successful applications. He described a particularly successful experiment in prediction of protein secondary structure, where not only was the ILP method better in terms of predictive accuracy than alternative published

methods, but, perhaps more significantly, yielded new domain knowledge. Still, much work remains to be done in handling noisy probabilistic concepts and especially in dealing with very large databases.

The workshop revealed that much work is afoot in the knowledge discovery area which promises to take us beyond the discovery of relatively simple representations such as conjunctive probabilistic rules or linear models. However, as one broadens the search space to allow for more expressive languages of knowledge representation, there comes an inevitable computational penalty in terms of scaling complexity of the algorithms. Each of the three talks showed that while the underlying models may be very different, for each class of models steady progress is being made on whittling down impractical algorithms to practical ones by taking advantage of particular structural characteristics of the methods and the representation being used. We hope to see some of the presented techniques showing up at future workshops as standard workhorses of successful applications.

Integrated and Interactive Systems

The third theme of the workshop dealt with Integrated and Interactive Systems. The two are closely related, since multi-method, integrated discovery systems frequently rely on human expertise to select the next discovery method, and interactive systems frequently offer a choice of multiple discovery algorithms.

Ron Brachman (AT&T Bell Laboratories) started the session with a talk about "Integrated Support for Data Archaeology", which is a skilled human task of interactive and iterative data segmentation and analysis. He presented a system, called IMACS, that supports a data archaeologist with a natural, object-oriented description of an application domain, a powerful query language, and a friendly user interface that, supports interactive exploration. IMACS is built on CLASSIC, a formal knowledge representation system.

Willi Kloesgen (GMD, Germany) described rule refinement and optimization strategies in Explora, an interactive system for discovery of interesting patterns in databases. The number of patterns presented to tile user is reduced by organizing the search hierarchically, beginning with the strongest, most general, hypotheses. An additional refinement strategy selects the most interesting statements and eliminates the overlapping findings. The efficiency of discovery is improved by inverting the record-oriented data structure and storing all values of the same variable together, which allows efficient computation of aggregate measures. Different data subsets are represented as bit-vectors making computation of logical combinations of conditions very efficient. Explora, a publicly available system,[1] which runs on a Mac, was demonstrated at tile workshop.

Philip Chan (Columbia U.) proposed Mcta-learning as a general technique to integrate a number of distinct learning processes. He examined several techniques of learning arbiters that select among independently learned classifiers. Such strategies are especially suitable for massive amounts of data that main-memory-based learning algorithms cannot efficiently handle. Preliminary results are encouraging, snowing that parallel learning by meta-learning can achieve comparable prediction accuracy in less time and space than purely serial learning.

An important design issue discussed at the workshop was the use of an internal vs. an external database. Both IMACS and Explora usc an *internal database* approach of pre-loading relevant parts of the data and transforming it into their internal and efficient format. his approach generally speeds up discovery for small or medium-size databases, However, it limits the system ability to work with large external databases. An *external database* approach, taken in discovery systems such as SKICAT, Spotlight, and KEFIR, is to build an interface, usually based on SQL, to a DBMS. This approach has its difficulties, such as dealing with communication problems and having to fit the discovery requests into the Procrustean bed of SQL. Retrieval from an external database may take longer, since in addition to a communication delay, the physical database organization may be sub-optimal for discovery system requests. However, this approach allows handling of large external databases that would not fit in memory and avoids duplicating the code for DBMS operations like joins or aggregations. We expect the coming advances in database technology, such as faster hardware, SQL servers, and forthcoming powerful SQL 2 and SQL 3 standards, to make the external database approach more attractive.

Other related issues were discussed at the summary session. Larry Kerschberg (George Mason U.)

[1] anonymous ftp to ftp.gmd.de or 129.26.8.90, ccl gmd/explora

observed that analysts frequently need to track hypotheses in multiple databases and proposed a mediator agent between an analyst and different discovery algorithms on one hand, and multiple data and knowledge source on the other hand. Mets-learning may offer a way to develop such mediator agents.

Jan Zytkow proposed an agenda for integration. The first part is integration of different forms of knowledge: contingency tables, rules, decision trees, and equations. Each form has different strengths and a limited conversion is possible from one form into another. The second part is integration of search in different spaces of new terms, equations, and rules. Such integration is required in a machine discovery system to match human flexibility in detecting patterns of different type. The multiple searches should be globally controlled and guided by a combination of data conditions, background knowledge, and user preferences.

Advances and Difficulties

The workshop and the following discussion on the KDD Nuggets e-mail list[2] highlighted several difficulties in application development.

Insufficient statistical **awareness**: Some KDD experiments arc performed without sufficient awareness of statistical theory. The classical example of this problem is testing N independent patterns for deviation from the norm, each test having a significance of cr. Then, $N\alpha$ patterns are likely to pass the test purely due to chance. Eliminating such "random" discoveries requires statistical controls, such as Bonferroni adjustments, which in the above example means reducing the significance level for each test to α/N, in order to assign the final discovery the significance of cr. Other ways to eliminate chance discoveries include randomized testing procedures (Jensen 1991). At the summary session John Major estimated that only shout half of the work presented at the workshop dealt adequately with this problem. Hopefully, raising this issue will increase proper statistical awareness.

Overabundance of **patterns**: As many pioneers of KDD have found, even with proper statistics, it is all too easy to find many statistically significant patterns which are either obvious, redundant, or useless. A common approach to reducing the number of obvious "discoveries" (such as only women have pregnancies), is to focus on changes, since "obvious" patterns will not change. Redundant discoveries can be eliminated by rule refinement methods such as those presented by Major or Kloesgen, or by using some findings to explain others. The more difficult task of separating the important patterns from the useless requires domain knowledge. A general heuristic here is that rules and patterns arc important to the degree they can lead to a useful action. This suggests a decision-theoretic framing of the problem of evaluating the usefulness of discovered patterns. The *utility* of a particular pattern should not be measured in isolation, but instead evaluated in tile context of set of possible actions.

Integration: Even if a perfect discovery systems is built, it needs to be integrated with other existing hardware/software systems to be useful. As expert system developers discovered years earlier, usually only a small part, of the deployed system is new technology - the rest is interfacing and system integration, mundane but critical steps in moving from prototype stage to deployment.

Privacy vs Discovery: Discovery in social or business data may raise a number of legal, ethical, and privacy issues. In 1990, Lotus was planning to introduce a CD-Rom with data on more than 100 million American households. The stormy protest led to the withdrawal of this product (Rosenberg 1992). Recent conferences on Computers, Freedom, and Privacy have also increased the awareness about issues of privacy and data ownership.

These difficulties are compensated by a number of important advances in areas relevant to KDD. Here we list only a few.

Multistrategy systems: Several recent comparisons of different learning and discovery algorithms have showed that different methods are superior for different types of problems (Brodley 1993) - no single method is best across a range of problems. As a result, there is a movement to multistrategy learning methods, especially for c].ossification, which apply a number of different methods to the *same* task and select rules from the best method. This is an area of very active research interest, with recent progress reported in (Michalski & Tecuci, 1993a, 1993 b).

[2] to subscribe, send e-mail to kdd@gte.com

Handling large scale databases: Since most learning algorithms cannot handle very large datasets, it is usually necessary to reduce the size of data on which learning is performed. One way is to eliminate irrelevant data using the data dependencies. This has been shown (Almuallim and Dietterich 1991) to increase the performance of the classifier methods. Other methods rely on various forms of data sampling. Catlett used an intelligent sampling approach to make a sublinear algorithm for decision tree induction (Catlett 1991). His method has been used to efficiently learn decision trees from databases with hundreds of thousands of records.

Overall, the workshop reflected measurable progress in developing and deploying KDD applications.

Acknowledgments: 'The authors thank all the workshop participants, and especially Inderpal Bhandari, Saso Dzeroski, Usama Fayyad, Willi Kloesgen, Ryszard Michalski, Evangelos Simoudis, and Jan Zytkow for their comments and contributions to this report. We thank Patel-Schneider for his editorial guidance and rapid processing of this paper.

References

H. Almuallim and T. Dietterich, 1991. Learning with Many Irrelevant Features. in Proceedings of AA A1-91, 547-552. Menlo Park, Calif: AAAI.

T. Anand and G. Kahn 1992. SPOTLIGHT: A Data Explanation System. Proceedings of CAIA-92. Washington, D.C.: IEEE Computer Society.

C. Brodley , 1993. Addressing the Selective Superiority Problem: Automatic Algorithm/Model Class Selection, In Proc. of 10th Machine Learning Conference, 17--24. Morgan Kaufmann.

J. Catlett, 1991. Megainduction: A Test Flight. in Proc. of 8th Machine Learning Conference, 596-599. Morgan Kaufmann.

N. Cercone, 1993. Guest editor, Special Issue on Learning and Discovery in Databases, *IEEE Trans. on Knowledge and Data Engineering, Dec.*

W. Frawley, G. I'iatetsky-Shapiro, and C. Matheus, 1992. Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992. Reprint of the introductory chapter of *Knowledge* Discovery *in Databases* collection, AAAI/MIT Press, 1991.

W. H. Inmon and S. Osterfelt, 1991. *Understanding Data Pattern Processing: the key to Competitive Advantage.* QED Technical Publishing Group, Wellesley, MA.

D. Jensen, 1991. Knowledge discovery through induction with randomization testing, in *Proceedings of the 1991 AAAI KDD Workshop, G.* Piatetsky-Shapiro (cd.), AAA1, Anaheim, CA, pp. 148-159.

N. Lavrač and S. Džeroski, 1993. *Inductive Logic Programming: Techniques and Applications.* Ellis Horwood, Chic. ıester.

C. Matheus, P. Chan, G. Piatetsky-Shapiro, 1993. Systems for Knowledge Discovery in Databases, *IEEE Trans. on Knowledge and Data Engineering,* Dec.

R. S. Michalski and G. Tecuci, 1993a. Editors, Proceedings of the 2nd International Workshop on Multistrategy Learning, George Mason University, Harpers Ferry, 1993.

R. S. Michalski and G. Tecuci, 1993b. Editors, *Machine Learning: A Multistrategy Approach, Volume IV.* Morgan Kaufmann.

S. Muggleton, 1992. *Inductive Logic Programming.* Academic Press, London.

K. Parsaye and M. Chignell, 1993. *Intelligent Database Tools & Applications. NY:* John Wiley.

J. l'earl, 1992. Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference, 2nd ed. San Mateo, Calif.: Morgan Kaufmann.

G. Piatetsky-Shapiro, 1991a. Knowledge Discovery in Real Databases: A workshop report on KDD-89, AI Magazine, vol. 11, no. 5, January 1991.

G. Piatetsky-Shapiro, 1991b. Report on AAAI-91 workshop on Knowledge Discovery in Databases, IEEE

Expert, October.

G. Piatetsky-Shapiro and W. Frawley, 1991. Editors, *Knowledge* Discovery *in Databases,* Cambridge, Mass.: AAAI/MIT Press.

G. Piatetsky-Shapiro, 1992. Editor, Special issue on Knowledge Discovery in Databases and KnowledgeBases, *Int. J. of Intelligent Systems* 7:7, *Sep.*

G. Piatetsky-Shapiro, 1993. Editor, Proceedings of KDD-93: the AAAI-93 workshop on Knowledge Discovery in Databases, AAAI Press report TR-20.

M. Rosenberg, 1992, Protecting Privacy, Inside Risks column, *Communications of ACM, 35(4),* p. 164.

P. Spirtes, C. Glymour, R. Scheines, 1993. *Causation, Prediction, and Search,* Lecture Notes in Statistics, Springer-Verlag.

W. Ziarko, 1993. Proceedings of the Rough Sets and Knowledge Discovery workshop, Banff, Canada.

J. Zytkow, 1992. Editor, Proceedings of the Machine Discovery Workshop, Aberdeen, Scotland, July.

J. Zytkow, 1993. Guest editor, Special Issue on Machine Discovery, *Machine Learning, 12(1-3).*