The Knowledge Engineering Review, Vol. 9:1, 1994, 57–60

Knowledge discovery in databases: Progress report

GREGORY PIATETSKY-SHAPIRO
GTE Laboratories Incorporated, 40 Sylvan Road, Waltham, MA 01254 USA (email: gps@gte.com)

1 Introduction

As the number and size of very large databases continues to grow rapidly, so does the need to make sense of them. This need is addressed by the field called Knowledge Discovery in Databases (KDD), which combines approaches from machine learning, statistics, intelligent databases, and knowledge acquisition. KDD encompasses a number of different discovery methods, such as clustering, data summarization, learning classification rules, finding dependency networks, analysing changes, and detecting anomalies (Matheus et al., 1993). Many applications of KDD have been reported in business, government and science (Piatetsky-Shapiro & Frawley, 1991; Parsaye and Chignell, 1993).

Over 60 researchers from ten countries took part in the third KDD workshop (Piatetsky-Shapiro, 1993) held at the American Association of Artificial Intelligence Conference in Washington, DC. A major trend evident at the workshop is the transition of research in the core KDD area (the discovery of relatively simple patterns in relational databases) to applications; with the most successful applications, such as SKICAT and Opportunity Explorer (described below), appearing in the areas of greatest need, where the databases are so large that manual analysis is impossible. Progress is 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 insufficient statistical awareness, overabundance of patterns, and lack of integration with existing systems.

This report focuses on the main themes of the workshop, which were also the topics of workshop sessions: Real-World applications, Discovery of Dependencies and Models, and Integrated and Interactive KDD Systems.

2 Real-world applications

KDD applications presented at the workshop were in diverse areas, including astronomy, insurance, marketing, software engineering, medicine, manufacturing, and stock market analysis.

Usama Fayyad (JPL) described SKICAT, an automated system for analysing large-scale, stellar sky surveys. Using a number of innovative machine learning methods, Usama and his colleagues were able to classify objects at least one magnitude fainter than in previous surveys and achieve an accuracy of about 94%. This work is one of the first applications of machine learning in actual use. In a related talk, Padhraic Smyth, also from JPL, 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.

Tej Anand presented A.C. Nielsen's recent work on a commercial product called Opportunity Explorer, an extension of their Spotlight product (Anand & Kahn, 1992) for identifying and reporting on trends and exceptional events in supermarket sales data. An important and innovative feature of Spotlight, which made it a best-selling product, was the automatic explanation of relationships between key events. Opportunity Explorer is a more general tool, suitable for different business sectors, for developing interactive, hypertextual reports using knowledge discovery templates which convert a large data space into concise, inter-linked information frames. Spotlight and Opportunity Explorer are in use by sales analysts and product managers of consumer packaged goods companies.

John Major (Travelers) dealt with the very important problem of selecting the most interesting rules among those discovered in data. He 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

are potentially redundant. In an application of the method to a tropical storm database, their system reduced 161 rules generated by IXL (a product of IntelligenceWare, Inc) to ten most interesting ones which were meaningful to a meteorologist.

Almost a dozen additional applications and system demonstrations were presented at the workshop's lively Posters and Demos session.

3 Discovery of dependencies and models

This was the second major theme of the workshop and the title of the third workshop session.

Jan Zytkow (Wichita State U.) outlined the latest developments in his research on deriving equations from data. He proposed a computationally simple test for the absence of functional dependency, in which case the much more expensive search to determine the form of dependency can be avoided. The test relies on discretization of data into smaller and smaller intervals.

Dependency networks are an important form of discovered knowledge, and recent progress in this field (Pearl, 1992) is very encouraging for KDD. Greg Cooper (U. of Pittsburgh), presented the latest results in his research on the use 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 algorithm.

Later, Cooper summarized the progress relevant in 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 are indistinguishable based only on data; new methods were developed for the discovery of probabilistic networks with measured and possibly unmeasured (latent) variables; these methods were applied to real data with promising results. The major improvements needed for applications to real databases are integrating different methods, and especially dealing with both discrete and continuous variables, improving search efficiency, and estimating the confidence and the stability of the output.

Saso Dzeroski (Josef Stefan Institute, Slovenia) presented an invited overview of Inductive Logic Programming (ILP) methods for KDD. ILP 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 general language of first-order relations. The field has developed rapidly in recent years (Muggleton, 1992), and now boasts relatively sophisticated algorithms and methods for handling a variety of problems, with a great potential for KDD applications. Dzeroski outlined 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 protein prediction, where ILP method not only had better predictive accuracy than alternative published methods, but perhaps more significantly, yielded new domain knowledge. Still, much work remains to be done; handling of noisy probabilistic concepts, for example, remains problematic for ILP methods in general.

The session indicated several directions in the knowledge discovery area which promise to take us beyond the discovery of relatively simple representations such as conjunctive probabilistic rules of linear models, into the more complex and more interesting representations such as dependency graphs and first-order relations. However, broadening the search space to allow more complex knowledge representation shows the search significantly, and makes the discovery at least NP-hard. However, progress is being made towards efficient approximate and heuristic algorithms.

4 Integrated and interactive systems

The last workshop session 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 step, and interactive systems frequently offer a choice of multiple discovery methods.

Ron Brachman (AT&T Bell Laboratories) talked 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 user with a natural, object-oriented description of an application domain, a powerful query language, and a friendly user interface for interactive exploration. IMACS is built on CLASSIC, a formal knowledge representation system.

Willi Kloesgen (GMD, Germany) described how Explora, an interactive system for discovery of interesting patterns in databases, does rule refinement. The amount of patterns presented to the 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.

Both IMACS and Explora pre-load the data into memory and transform it into their internal representation. While this approach speeds up discovery, it limits their system's ability to work with external or very large databases. A different approach for discovery system is to build an interface to external DBMS. This allows handling of very large external databases and avoids duplicating the code for DBMS operations.

Philip Chan (Columbia) proposed Meta-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, showing that parallel learning by meta-learning can achieve comparable prediction accuracy in less time and space than purely serial learning.

5 Difficulties to overcome

A number of difficulties remain in the path of developing and deploying KDD applications.

- *Insufficient statistical awareness:* The classical example is if 100 independent random variables are tested for deviation from the norm with a significance of 0.01, one of them is likely to pass the test purely due to chance. This example is very relevant to KDD, since the number of potential rules or patterns is typically exponential in the number of data fields and there is a danger of making "random" discoveries. This problem exists in quite a number of current KDD systems. The solution is using proper statistical controls, such as Bonferroni adjustments.
- *Obvious, redundant and useless patterns:* As many pioneers of KDD have found, even with proper statistics, systems still find too many patterns which are either obvious, redundant, or useless to the user. A common approach to reducing the number of obvious "discoveries" (such as only women have pregnancies) is to focus on changes, since "obvious" patterns do not change. Redundant discoveries can be dealt with by rule refinement methods which eliminate rules similar to each other, or by using some findings to explain others. The more difficult task of eliminating useless patterns requires domain knowledge. A general heuristic here is that rules and patterns are important to the degree they can lead to a useful action.
- *Integration with existing systems:* To be useful, even a perfect discovery system needs to be integrated with other existing systems, such as DBMSs, spreadsheets, statistical and graphical packages, etc. As expert system developers discovered years earlier, usually only a small part of the successful 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. Recent conferences on Computers, Freedom and Privacy have also increased awareness about the issues of privacy and data ownership.

Despite these difficulties, the workshop reflected measurable progress in developing and deploying KDD applications.

Acknowledgements

I am grateful to Chris Matheus, Padhraic Smyth, Sam Uthurnsamy, and Bud Frawley for their contributions to this report.

References

Anand, T and Kahn, G, 1992. "SPOTLIGHT: A data explanation system". In: *Proceedings of CAIA-92*. Washington, DC: IEEE Computer Society.

Matheus, C, Chan, P and Piatetsky-Shapiro, G, 1993. "Systems for knowledge discovery in databases". *IEEE Transactions on Knowledge and Data Engineering*. December.

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

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

Pearl, J, 1992. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference, 2nd edition*, San Mateo, CA: Morgan Kauffman.

Piatetsky-Shapiro, G and Frawley, W (eds), 1991. *Knowledge Discovery in Databases*, Cambridge, MA: AAAI/MIT Press.

Piatetsky-Shapiro, G, (ed.), 1993. *Proceedings of KDD-93: The AAAI-93 Workshop on Knowledge Discovery in Databases*. Menlo Park, CA: AAAI Press.