

Knowledge Mining in Databases: An Integration of Machine Learning Methodologies with Database Technologies*

Jiawei Han, Yongjian Fu, Krzysztof Koperski, Gabor Melli, Wei Wang, Osmar R. Zaiane

Knowledge Discovery Research Group
Database Systems Research Laboratory
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada V5A 1S6

E-mail: {han, yongjian, koperski, melli, weiw, zaiane}@cs.sfu.ca

Abstract

Active research has been conducted on knowledge discovery in databases by the researchers in our group for years, with many interesting results published and a prototyped knowledge discovery system, DBMiner (previously called DBLearn), developed and demonstrated in several conferences. Our research covers a wide spectrum of knowledge discovery, including (1) the study of knowledge discovery in relational, object-oriented, deductive, spatial, and active databases, and global information systems, and (2) the development of various kinds of knowledge discovery methods, including attribute-oriented induction, progressive deepening for mining multiple-level rules, meta-rule guided knowledge mining, etc. Techniques for the discovery of various kinds of knowledge, including generalization, characterization, discrimination, association, classification, clustering, etc. and the application of knowledge discovery for intelligent query answering, multiple-layered database construction, etc. have also been studied in our research.

1 Introduction

With the rapid growth of the number of databases and the tremendous amounts of data being collected and stored in databases, it is increasingly important to develop software tools to assist in the extraction of “information” or “knowledge” from data, understanding the implications of data in databases, and automatic construction of knowledge-bases from databases. The research into *knowledge discovery in databases* (or *data mining*) [1, 18] has attracted wide attention in both academia and industry.

The Knowledge Discovery research group in the Database Systems Research Laboratory of the School of Computing Science, Simon Fraser University, has been working in this promising research field for several years and has contributed to the field in the following aspects.

1. The development of a prototyped knowledge discovery system, DBMiner (previously named DBLearn) [8, 10], which integrates machine learning methodologies with database technologies and discovers different kinds of knowledge from large databases efficiently and effectively. The system may discover different kinds of knowledge, including characteristic, discriminant, association, and classification rules using a set of knowledge discovery methods, originated from our own research, including attribute-oriented induction [3], progressive deepening for mining multiple-level rules [6], meta-rule guided knowledge mining [2], etc.
2. The study of knowledge discovery in different kinds of databases [14, 12, 17, 13, 20], including knowledge discovery in relational, object-oriented, deductive, spatial, and active databases, and global information systems, and the application of knowledge discovery for intelligent query answering [11], multiple-layered database construction [9], etc.

*Research is partially supported by the Natural Sciences and Engineering Research Council of Canada under the grant OGP0037230, by the Networks of Centres of Excellence Program (with the participation of PRECARN association) under the grants IRIS:HMI-5 and IRIS:IC-2, and by a research grant from the Hughes Research Laboratories.

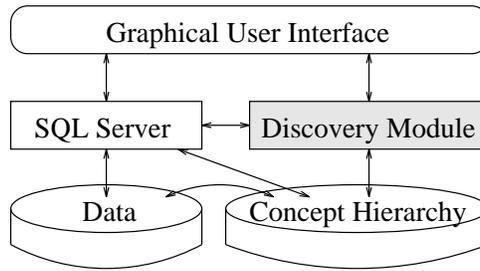


Figure 1: General architecture of DBMiner

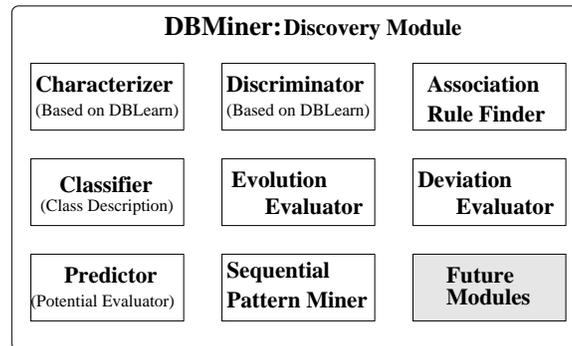


Figure 2: Function modules of DBMiner

The remaining of the paper is organized as follows. A more detailed description of the DBMiner system is presented in Section 2. Knowledge mining in advanced database systems and knowledge discovery applications are outlined in Section 3. A summary and a discussion of our on-going research is presented in Section 4.

2 DBMiner: A database mining system

DBMiner, a comprehensive database mining system prototype, has been developed in Simon Fraser University. It started with an interesting induction method, attribute-oriented induction [3, 4], for learning characteristic rules and discriminant rules in relational databases. The method resulted in an early version of the system, DBLearn [4, 8]. Experiments with DBLearn were performed in NSERC (Natural Science and Engineering Research Council of Canada) research grant information database and in several large industrial databases with successful results and good performance. Further extensions and enhancements of the DBLearn system since 1993 have led to a new generation of the system: DBMiner [7]. DBMiner consists of several new functional modules besides the characterizer and discriminator in DBLearn. It performs dynamic adjustment of concept hierarchies and automatic generation of numeric hierarchies. It discovers different kinds of knowledge rules and generates different forms of outputs, including generalized relations, generalized feature tables, and multiple forms of generalized rules. Moreover, system performance has been improved, graphical user interfaces have been enhanced for interactive knowledge mining, and a client/server architecture has been constructed for industrial applications.

The major features of the system include the integration of machine learning and database technologies, high speed and efficiency in analyzing large databases, interactive knowledge mining, and smooth intergration with commercial relational database systems.

Figure 1 shows the general architecture of DBMiner which integrates a relational database system, such as a Sybase/Oracle SQL server, with the discovery module. The core of the DBMiner system is the discovery module, which is further detailed in Figure 2. It consists of multiple functional modules, including characterizer, discriminator, association rule finder, classifier, evolution evaluator, etc.

The functionalities of the first three modules are described as follows:

- The characterizer [4] discovers a set of characteristic rules from the relevant set of data in a database. A characteristic rule summarizes the general characteristics of a set of user-specified data.

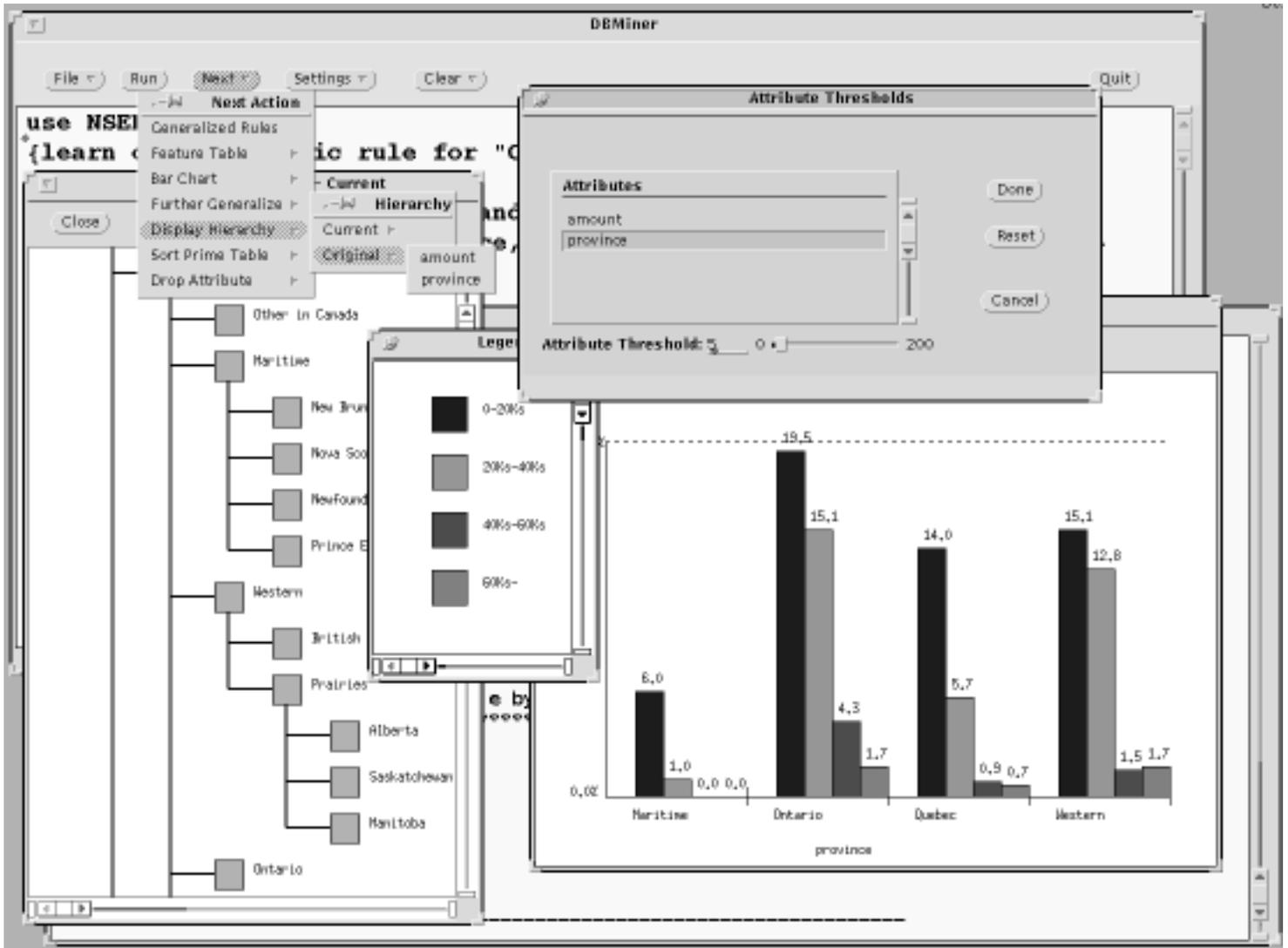


Figure 3: A snapshot of the execution of DBMiner

- A discriminator [4] discovers a set of discriminant rules from the relevant set(s) of data in a database. A discriminant rule distinguishes the general features of one set of data, called the *target class*, from some other set(s) of data, called the *contrasting class(es)*.
- An association rule finder [6] discovers a set of association rules (in the form of “ $A_1 \wedge \dots \wedge A_i \rightarrow B_1 \wedge \dots \wedge B_j$ ”) at multiple concept levels from the relevant set(s) of data in a database. For example, it may find from a large set of transaction data an association rule, such as if a customer buys (*one brand of*) milk, s/he usually buys (*another brand of*) bread.

In the development of other functional modules, attribute-oriented induction [4, 7] also plays an essential role. It integrates a machine learning paradigm *learning-from-examples* [15] with set-oriented database operations and substantially reduces the computational complexity of database learning processes. Moreover, the generalized relation can be further analyzed by integration with other machine learning methods [7], including ID-3 [19], Cluster-2 [16], etc.

The system also performs *automatic generation of conceptual hierarchies* for numerical attributes and *dynamic conceptual hierarchy adjustment* [5] for all the attributes based on the statistical distribution of the set of relevant data, which produces desirable generalized results.

DBMiner offers both graphical and SQL-like interfaces [7]. For example, to characterize *Computer Science* grants in the *NSERC94* database in relevance to discipline and amount categories and the distribution of count% and amount%, the data mining query is as follows.

```
use NSERC94
characterize "CS_Discipline_Grants"
from award A, grant_type G
where A.grant_code = G.grant_code and A.disc_code = "Computer"
in relevance to disc_code, amount, percentage(count), percentage(amount)
```

To process this query, the system first obtains the relevant set of data by processing a relational database query, then generalizes the data using an attribute-oriented induction approach [3, 4], and presents different forms of outputs. The output outlines the number or amount distribution of computer science (research) grants according to discipline categories (such as *theory*, *AI*, *database*, and so on). The output forms include generalized relations, generalized feature tables, bar charts, and generalized rules.

A snapshot of the execution of DBMiner is presented in Figure 3. The DBMiner system is accessible with the Internet address: <http://www.dbg.sfu.ca/dbl/dbminer> or <http://fas.sfu.ca/cs/research/groups/DB/dbminer>.

3 Knowledge discovery in advanced database systems and knowledge discovery applications

Beside knowledge discovery in large relational databases, investigations have also been performed on efficient and effective methods for knowledge discovery in object-oriented databases [12], spatial databases [13, 14, 17], active databases [12], deductive databases [2], transaction databases [6], and global information systems [20]. Three of them are outlined below to convey the ideas.

- For knowledge discovery in object-oriented databases [12], techniques have been studied on generalization of complex objects, attributes and methods, class and aggregation hierarchies, spatial and multimedia data, etc. After generalization of complex objects and structures into relational-like data, most techniques developed for data mining in relational systems can be applied for knowledge mining in object-oriented databases.
- For knowledge discovery in spatial databases [13, 14, 17], a set of techniques have been developed in our research, including nonspatial-data-dominant generalization [14], spatial-data-dominant generalization [14], spatial data clustering [17], and spatial feature association [13].
- For resource and knowledge discovery in global information systems (Internet) [20], a multiple-layered database structure has been proposed, which first transforms the low-level, highly unstructured, widely distributed, heterogeneous global information-base into relatively structured higher layer databases by various kinds of generalization techniques and then constructs multiple-layered information-base to facilitate resource and knowledge discovery. Preliminary experiments on a small subset of networked information-base have demonstrated some limited success of the approach [20].

Knowledge discovery has strong application potential. Besides the use of discovered knowledge for decision making, process control and knowledge-base construction, we have investigated its application in intelligent query answering [11] and multiple-layered database construction [9]. Database queries can be answered intelligently by analyzing the intent of query and providing generalized, neighborhood or associated information using stored or discovered knowledge. Knowledge discovery substantially broadens the spectrum of intelligent query answering and provides discovered knowledge and knowledge discovery tools, which include generalization, data summarization, concept clustering, rule discovery, query rewriting, deduction, lazy evaluation, construction and application of multiple-layered databases, etc.

4 A summary of our on-going work

Our research progress on knowledge discovery in databases have greatly motivated our on-going work in this direction. These on-going studies are illustrated as follows.

- Further enhancement of the power and efficiency of the knowledge discovery mechanisms [7], including the improvement of rule quality and system performance for the existing functional modules, the development of techniques for mining new kinds of rules, etc.
- Further development of high-level, user-friendly interfaces for interactive knowledge mining. This includes multiple platforms of knowledge mining interfaces, including X-window-oriented, PC-window-oriented, and Netscape(WWW)-oriented interfaces, and visual presentation of the discovered knowledge.
- Extension of data mining technique to advanced and/or special purpose database systems, including further studies on spatial data mining, knowledge mining in heterogeneous databases, and knowledge discovery in global information systems. This may lead to the generation of a set of new, special-purpose data mining systems, such as GeoMiner, WebMiner, etc.

References

- [1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1995.
- [2] Y. Fu and J. Han. Meta-rule-guided mining of association rules in relational databases. In *Proc. 1st Int'l Workshop on Integration of Knowledge Discovery with Deductive and Object-Oriented Databases (KDOOD'95)*, Singapore, Dec. 1995.
- [3] J. Han, Y. Cai, and N. Cercone. Knowledge discovery in databases: An attribute-oriented approach. In *Proc. 18th Int. Conf. Very Large Data Bases*, pages 547–559, Vancouver, Canada, August 1992.
- [4] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 5:29–40, 1993.
- [5] J. Han and Y. Fu. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In *Proc. AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94)*, pages 157–168, Seattle, WA, July 1994.
- [6] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. 1995 Int. Conf. Very Large Data Bases*, Zurich, Switzerland, Sept. 1995.
- [7] J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1995.
- [8] J. Han, Y. Fu, Y. Huang, Y. Cai, and N. Cercone. DBLearn: A system prototype for knowledge discovery in relational databases. In *Proc. 1994 ACM-SIGMOD Conf. Management of Data*, page 516, Minneapolis, MN, May 1994.
- [9] J. Han, Y. Fu, and R. Ng. Cooperative query answering using multiple-layered databases. In *Proc. 2nd Int. Conf. Cooperative Information Systems*, pages 47–58, Toronto, Canada, May 1994.
- [10] J. Han, Y. Fu, and S. Tang. Advances of the DBLearn system for knowledge discovery in large databases. In *Proc. 1995 Int'l Joint Conf. on Artificial Intelligence*, Montreal, Canada, Aug. 1995.
- [11] J. Han, Y. Huang, N. Cercone, and Y. Fu. Intelligent query answering by knowledge discovery techniques. In *IEEE Trans. Knowledge and Data Engineering (accepted)*, 1995.
- [12] J. Han, S. Nishio, and H. Kawano. Knowledge discovery in object-oriented and active databases. In F. Fuchi and T. Yokoi, editors, *Knowledge Building and Knowledge Sharing*, pages 221–230. Ohmsha, Ltd. and IOS Press, 1994.
- [13] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int'l Symp. on Large Spatial Databases (SSD'95)*, Portland, Maine, Aug. 1995.
- [14] W. Lu, J. Han, and B. C. Ooi. Knowledge discovery in large spatial databases. In *Far East Workshop on Geographic Information Systems*, pages 275–289, Singapore, June 1993.
- [15] R. S. Michalski. A theory and methodology of inductive learning. In Michalski et al., editor, *Machine Learning: An Artificial Intelligence Approach, Vol. 1*, pages 83–134. Morgan Kaufmann, 1983.
- [16] R. S. Michalski and R. Stepp. Automated construction of classifications: Conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 5:396–410, 1983.
- [17] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases*, pages 144–155, Santiago, Chile, September 1994.
- [18] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
- [19] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1992.
- [20] O. R. Zaïane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. In *Proc. 1st Int'l Conf. on Knowledge Discovery and Data Mining (KDD'95)*, Montreal, Canada, Aug. 1995.

About the authors

- Jiawei Han. Professor, Computing Science, Simon Fraser University. His major research interests include database and knowledge-base systems, knowledge discovery in databases, deductive and object-oriented databases, spatial and multi-media databases, logic programming, and artificial intelligence. He is the leader of the Canadian IRIS project IRIS:HMI-5 (“Data Mining and Knowledge Discovery in Large Databases”), and has served or is currently serving in the program committees of over 20 international conferences, including ICDE’95 (also, program committee vice-chairman), DOOD’95, ACM-SIGMOD’96, and VLDB’96.
- Yongjian Fu. Ph.D. student, Computing Science, Simon Fraser University. He received M.Sc. and B.Sc. degrees at Zhejiang University, China, and worked as a lecturer there before joining SFU. He is the major developer of the DBMiner system and works on research into knowledge discovery from databases and applications of knowledge discovery systems.
- Krzysztof Koperski. Ph.D. student, Computing Science, Simon Fraser University. His major research focus is on spatial data mining. He is also interested in spatial reasoning and spatial object-oriented databases.
- Gabor Melli. M.Sc. student, Computing Science, Simon Fraser University. He is an expert on systems and has contributed to the system support of the relational database system engine. His current research focuses on automatic prediction of attribute-value behavior.
- Wei Wang. M.Sc. student, Computing Science, Simon Fraser University. He has been implementing the GUI interface for the DBMiner system on PCs and is performing research on knowledge discovery in heterogeneous databases.
- Osmar R. Zaiane. Ph.D. student, Computing Science, Simon Fraser University. He holds an M.Sc. degree in Computer Science from Laval University (Quebec, Canada) where he worked on mobile databases stored on smart-cards. He also holds an M.Sc. degree in Electronics from Paris XI University (Paris, France) where he worked on Natural Language Interfaces for textual databases. His current research focuses on knowledge discovery in global network information systems.