Advances of the DBLearn System for Knowledge Discovery in Large Databases*

Jiawei Han **Yongjian Fu** Simon **Tang**
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada V5A 1S6

Abstract

A prototyped data mining system, DBLearn, was developed in Simon Fraser Univ., which integrates machine learning methodologies with database technologies and efficiently and effectively extracts characteristic and discriminant rules from relational databases. Further developments, of DBLearn lead to a new generation data mining system: DBMiner, with the following features: (1) mining new kinds of rules from large databases, including multiple-level association rules, classification rules, cluster description rules, etc., (2) automatic generation and refinement of concept hierarchies, (3) high level SQL-like and graphical data mining interfaces, and (4) client/server architecture and performance improvements for large applications. The major features of the system are demonstrated with experiments in a research grant information database.

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 for *data mining* or *knowledge discovery in databases* [Piatetsky-Shapiro and Frawley, 1991; Fayyad *et ai,* 1995].

Data mining is the extraction of "information" or "knowledge" from data, which helps understanding data in databases and automatic construction of knowledge-bases from databases.

DBLearn is such a knowledge discovery system prototype, developed in Simon Fraser University between 1989 and 1993 [Cai *et* a/., 1991; Han *et* a/., 1993; 1994]. It discovers characteristic rules and discriminant rules embedded in relational databases. The major features of the system are speed and efficiency in analyzing large databases, interactive knowledge mining, and smooth integration with commercial relational database systems. Experiments with DBLearn have been 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.

*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 grant IRIS:HMI-5, and by a research grant from the Hughes Research Laboratories.

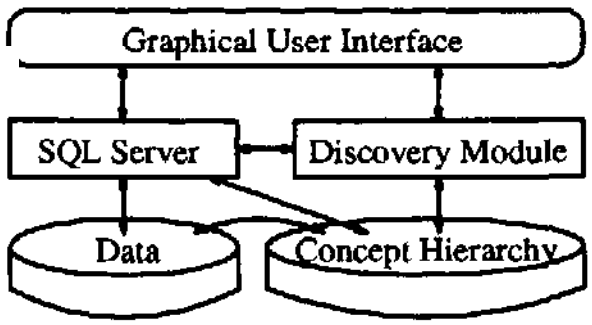

Figure 1. General architecture of DBMiner

Further extensions and enhancements of the DBLearn system since 1993 have led to a new generation of the system: DBMiner [Han and Fu, 1995]. 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 generates different forms of knowledge, 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.

2 Architecture and functionalities

The general architecture of DBMiner is shown in Figure 1 which tightly integrates a relational database system, such as a SyBase SQL server, with the discovery module. The discovery module of DBMiner shown in Figure 2 consists of multiple functional modules, including characterizer, discriminator, association rule finder, classifier, evolution evaluator, deviation evaluator, predictor, sequential pattern miner, and future modules. This video demonstrates the functionalities of the first three modules which are described as follows.

- The characterizer 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.

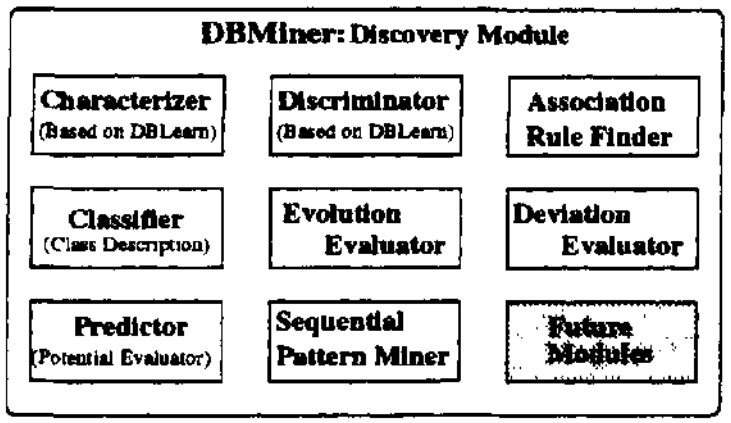

Figure 2: Function modules of DBMiner

- A discriminator 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 discovers a set of association rules (in the form of *"A_1* A ■ ■ • A *A_1* —► B_1 A ■ ■ ■ A B/[1]) 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 that if a customer buys (one *brand of)* milk,s/he usually buys *(another brand of)* bread.

DBMiner offers both graphical and SQL-like interfaces. 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.disccode = "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 [Cai *et ai,* 1991; Han et *al.,* 1993], and presents different forms of outputs, including generalized relations, generalized feature tables, bar/pie charts, generalized rules, to outline the number or amount distribution of computer science (research) grants according to discipline categories (such as *theory, AI, database,* and so on),

In the development of other functional modules, attribute-oriented induction [Han *et al,* 1993; Han and Fu, 1995] also plays an essential role. It integrates a machine learning paradigm *learning-from-examples* [Michalski, 1983] with set-oriented database operations and substantially reduces the computational complexity of database learning processes.

The system also performs *automatic generation of conceptual hierarchies* for numerical attributes and *dynamic conceptual hierarchy adjustment* [Han and Fu, 1994] for all the attributes based on the statistical distribution of the set of relevant data, which produces desirable generalized results.

3 Further development of DBMiner

The DBMiner system is currently being extended in several directions as follows.

- Further enhancement of the discovery power and efficiency for data mining in relational systems [Han and Fu, 1995], including the improvement of rule quality and system performance for the existing functional module, the development of techniques for mining new kinds of rules, etc.
- Integration, maintenance and application of discovered knowledge, including incremental update of discovered rules, merging of discovered rules into existing knowledge-bases, intelligent query answering using discovered knowledge, and the construction of multiple layered databases.
- Extension of data mining technique towards advanced and/or special purpose database systems, including extended-relational, object-oriented, deductive, spatial, temporal, and heterogeneous databases.

References

[Cai *et al.,* 1991] Y. Cai, N, Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley (eds.), *Knowledge Discovery m Databases,* pages 213-228. AAA1/MIT Press, 1991.

[Fayyad *et al,* 1995] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances* in *Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1995.

[Han *et al,* 1993] 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.

[Han and Fu, 1994] J. Han and Y. Fu. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In *AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94),* pages 157-168, Seattle, WA, July 1994.

[Han and Fu, 1995] J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, et al. (eds.), *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1995.

[Han *et al,* 1994] 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.

[Michalski, 1983] R. S. Michalski. A theory and methodology of inductive learning. In Michalski et al. (eds.), *Machine Learning: An Artificial Intelligence Approach, Vol. 1,* pages 83-134. Morgan Kaufmann, 1983.

[Piatetsky-Shapiro and Frawley, 199l] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases.* AAAI/MIT Press, 1991.