

DBLearn: A System Prototype for Knowledge Discovery in Relational Databases *

Jiawei Han [†] Yongjian Fu [†] Yue Huang [†] Yandong Cai [†] Nick Cercone ^{‡†}

[†] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

[‡] Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

Abstract

A prototyped data mining system, DBLearn, has been developed, which efficiently and effectively extracts different kinds of knowledge rules from relational databases. It has the following features: high level learning interfaces, tightly integrated with commercial relational database systems, automatic refinement of concept hierarchies, efficient discovery algorithms and good performance. Substantial extensions of its knowledge discovery power towards knowledge mining in object-oriented, deductive and spatial databases are under research and development.

Introduction

Research and development activities have been upsurging recently on *knowledge discovery in database (KDD)* for mining knowledge from large databases [4, 5]. A prototyped KDD system, DBLearn, has been developed based on our research [1, 2, 3], with the following features:

- The system adopts an *attribute-oriented induction* approach [1, 2, 3] which integrates a machine learning paradigm *learning-from-examples* with set-oriented database operations and *substantially reduces the computational complexity* of database learning processes.
- The system efficiently and effectively extracts different kinds of generalized rules, including *characteristic rules, discriminant rules, data classification rules, etc.*, from large relational databases.

*This work was supported in part by the research grants from *Natural Sciences and Engineering Research Council (NSERC) of Canada* and from the *Centre for Systems Science of Simon Fraser University*.

- The system is *tightly integrated with commercial relational database systems*. It uses an *SQL-like learning interface* and a *graphics user interface* and takes a relational system (e.g., SyBase) as its search engine.
- The system can perform *automatic generation of conceptual hierarchies for numerical attributes* and *dynamic conceptual hierarchy adjustment* based on the statistical distribution of the set of relevant data.
- Experiments have been conducted on several relational databases, including *Canada NSERC research grant information system, etc.* with impressive learning results and satisfactory performance.

Furthermore, the project will be extended substantially under the new IRIS-2 project (funded by Canadian *Networks of Centre of Excellence* Program), with a joint effort by a group of researchers from four Canadian universities and some industrial firms, towards efficient knowledge mining in extended-relational, object-oriented, deductive, spatial and heterogeneous databases, intelligent querying of data & knowledge, and control of dynamic processes.

References

- [1] Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 213–228. AAAI/MIT Press, 1991.
- [2] 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.
- [3] 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.
- [4] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
- [5] A. Silberschatz, M. Stonebraker, and J. D. Ullman. Database systems: Achievements and opportunities. *Comm. ACM*, 34:94–109, 1991.