

# Data Mining and Database Systems: Where is the Intersection?

Author: Surajit Chaudhuri

Ramesh Radhakrishnan

# Introduction

- What role, if any, does database research contribute to the area of data mining?

## Advantage of data mining over OLAP

- data mining tools automatically discover interesting patterns
  - such functionality will be useful in enterprise databases
- Important views presented in this paper:
    - Need to focus on generic scalability requirements
    - Need to build data mining systems that are “SQL-aware”

# Data mining landscape

Important work in data mining can be characterized as:

- Inventing new data analysis techniques
  - requires insight into statistical and machine learning and related algorithmic areas
  - little interaction with database system issues
- Scaling data analysis techniques over large data sets
  - large number of data records not been studied, unlike studies where number of dimensions is large
  - assumptions about data distribution that model a data set
  - multi-level store

# Scaling data analysis techniques

- Develop efficient algorithms that take into account the large data set
  - algorithms that carefully stage computations
  - Things to watch out for:
    - restricting the choice of data mining sets
    - scaling specific algorithms
    - ignoring sampling as a scaling methodology
- Restrict the scope of analysis objectives
  - strike a balance between accuracy and exhaustiveness of analysis with desire to be efficient
  - use support and confidence parameters. Guarantee of analysis can be probabilistic.

# Motivation for SQL-aware data mining systems

- Data is in the warehouse
  - which deploy relational database technology for storing and maintaining data
  - data will be shared by OLAP and other database utilities
- SQL systems can be leveraged
  - SQL DBMS provide a rich set of primitives for data retrieval
- Ad-hoc mining
  - algorithms at present are invoked on disk-resident data set
  - mining is invoked on data set that is created on the fly by query tools
  - exploit interaction of mining operators with SQL operators

# Building SQL-aware data mining systems

- Implementation of a classical main memory implementation of a decision tree classifier and Microsoft SQL server at Microsoft research
- Using SQL backend
  - could hurt performance
  - exploit physical database design and query processing subsystem
  - novel ways of staging computation
- SQL Extensions
  - strongly interact with core SQL primitives and improve performance
  - encapsulate a set of useful data mining primitives

# Summary

- Reviewed various facets of data mining
- generic scalability extensions for each class of data mining algorithms
- scalable implementations over SQL systems
- core data mining algorithms need to be integrated with other database tools