Theoretical Frameworks for Data Mining

(Heikki Mannila, 2000)

Presented by:

Noamaan Nadeem

2004-02-0133

Theory …

- Why look for it ?
- What purpose should it achieve ?

Theory should …

- Be simple and easy to apply.
- Lead to useful results.
- Model typical data-mining tasks.
- Model the probabilistic nature of discovered patterns.
- Encompass various forms of data.
- Capture various characteristics of data-mining.

Statistics

Differences

- Volume of data
- Number of variables / attributes
  - Impacts analysis method.
  - Computational feasibility an issue here.
- Data mining is secondary data analysis (David Hand).

Differences (cont…)

- Emphasis on database integration, simplicity of use, understandability of results in data mining.
- Statistics doesn’t look at the ‘process’ in data mining.

Applied Machine Learning

- Doesn’t encompass the statistics in data mining.
- Theoretical machine learning approaches do not address requirements for data mining theory.

Probabilistic Approach

- Finding underlying joint distributions between variables of data.
- Bayesian Model fits.

Data Compression

- Goal: compress data set by finding some structure for it.
- Can use Minimum Description Length (MDL) framework.
- Linked to Bayesian approach
- Framework doesn’t capture process view.

Microeconomic View

- Data mining is used to find actionable patterns (that increase utility).
- Can use theoretical formulation by Klein berg et. al.
- Captures pattern discovery, clustering.

Inductive Databases

- Apply the query concept: *there is no such thing as discovery; all is power of query language.*
- Uncover previously unseen information.
- Linked to theory of deductive databases.
- Captures the ‘process’ view of data mining.
- Fits in association rules, simple pattern formalism.
- Relies on probabilistic approach for clustering / probabilistic nature.

Further work…

- Heikki.mannila@nokia.com
- J. Kleinberg, C. Papadimitriou, and P. Raghavan, A Microeconomic View of Data Mining, *Data Mining and Knowledge Discovery* 2, 4 (1998), 311-324.
- A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin: Bayesian Data Analysis, *Chapman & Hall* 1995.