What is Data Mining ?

1 Overview

Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (from D.Hand, H. Mannila, P. Smyth, *Principles of Data Mining*)

Observational Data Sets

A data set (or data collection) is *observational* when it had been collected by its owner over a natural course of events.

A data set (or data collection) is *experimental* when its collection had been specifically set up and supervised by the owner.

E.g.: database of bank transactions (observational) vs. dataset from the survey of bank customers (experimental).

Key difference: Objectives of data mining tasks play no role in date collection.

Key issue: **Huge data sets !!!**

Unsuspected Relationships

Data Mining searches of existing *structure* in the data set.

Two types of structure:

Model is a comprehensive summary of the data collection. Models "make statements" about every point in the data set.

Pattern is a noted relationship that holds for part of the data set.

For example compare classifying supermarket shopping trips into the categories of "weekly grocery shopping", "quick stop", "party preparation" and "misc" (model) vs., noticing that most people who buy eggs and bread, also buy milk (pattern).

to Summarize the Data in Novel Ways

Discovered relationships (patterns and models) have to be communicated to the data set owner.

Problems are:

- size of the data set under study;
- size of the obtained model or patterns;
- multi-dimensionality of data.

Issues are:

- representation of relationships;
- selection of relationships that are general;
- visualization of results of data mining.

Useful to the Data Owner

Implicit here, is the assumption that data mining methods must work efficiently, i.e., the data owner, must *eventually* see the output (and *eventually* better happen sooner rather than later).

2 Three Sources and Four Components of Data Mining

2.1 Three Sources

2.1.1 Statistics

Statistics is the study of patterns (or correlations) in data.

Why Data Mining is NOT just Statistics ?

- *Statistics*: For statistical analysis of a phenomenon, first, the predictive model is built, then, the data is collected, and then, the fit of the model to the data is determined.

  I.e., *model* is the starting point.
- *Data Mining*: First the data is collected, then the data mining techniques are applied to produce the model of the data collection.

  I.e., *model* is the end point.

In Statistics, process similar to Data Mining is called *Exploratory Data Analysis*.

Why Data Mining is NOT just Exploratory Data Analysis ?

Key Issue: Size of the data set. Because the data sets are **HUGE**, traditional statistical methods may seize being efficient.

Cue in...

2.1.2 Artificial Intelligence (AI)

AI has many definitions, but most boil down to AI being the study of problems that have huge solution spaces and, therefore, require "intelligent" strategies for searching for good solutions.

Learning or *Machine Learning* is a field of AI that deals with construction of models for complex and large domains based on information "learned" from data.

From an AI point of view, *data mining* is one of the applications of machine learning methods.

2.1.3 Databases

...Did we mention that the datasets are HUGE and the methods must be *efficient* ?

Issues are:

- Organization of data sets on primary/secondary/tretiary storage;
- Indexing and efficient access to the data;
- Data manupulation.

All this is collectively known as *data management strategy.*

2.2 Four Components

1. *Model/Pattern structure.* What are we looking for ? (Statistics/AI)
2. *Score Function.* How do we tell success ? (Statistics)
3. *Search Method.* How do we look for the answer ? (AI)
4. *Data Management Strategy.* How do we efficently support high-level data mining tasks ? (Databases)