

Data mining on very large databases



Elena Baralis, Tania Cerquitelli

Politecnico di Torino

<http://dbdmg.polito.it/wiki/Diri/View/Public/WorkHome>

Torino, may 13th 2009



Outline

- Data mining fundamentals
- Association rules fundamentals
- Disk-based pattern mining
- Other research topics



2

Data mining fundamentals



Elena Baralis, Tania Cerquitelli

Politecnico di Torino



Data analysis

- Most companies own huge databases containing
 - operational data
 - textual documents
 - experiment results
- These databases are a potential source of useful information

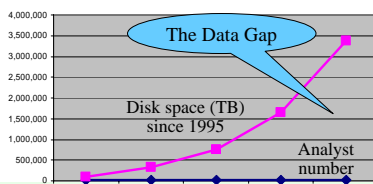


4



Data analysis

- Information is "hidden" in huge datasets
 - not immediately evident
 - human analysts need a large amount of time for the analysis
 - most data *is never analyzed at all*



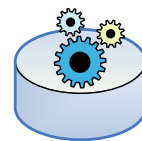
From R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

5



Data mining


- Non trivial extraction of
 - implicit
 - previously unknown
 - potentially useful information from available data
- Extraction is automatic
 - performed by appropriate algorithms
- Extracted information is represented by means of abstract models
 - denoted as *pattern*





6

Example: biological data

- Microarray
 - expression level of genes in a cellular tissue
 - various types (mRNA, DNA)
- Patient clinical records
 - personal and demographic data
 - exam results
- Textual data in public collections
 - heterogeneous formats, different objectives
 - scientific literature (PubMed)
 - ontologies (Gene Ontology)



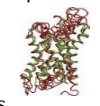

| CLL | PATIENT | probe1 | probe2 | probe3 | probe4 | probe5 | probe6 | probe7 | probe8 | probe9 | probe10 |
|-----------------|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| RANGE: 0.000000 | 0.00 | 0.20 | 0.40 | 0.60 | 0.80 | 1.00 | 1.20 | 1.40 | 1.60 | 1.80 | 2.00 |
| RANGE: 0.000000 | 0.00 | 0.20 | 0.40 | 0.60 | 0.80 | 1.00 | 1.20 | 1.40 | 1.60 | 1.80 | 2.00 |
| RANGE: 0.000000 | 0.00 | 0.20 | 0.40 | 0.60 | 0.80 | 1.00 | 1.20 | 1.40 | 1.60 | 1.80 | 2.00 |

DBGM 7

Biological analysis objectives

- Clinical analysis
 - detecting the causes of a pathology
 - monitoring the effect of a therapy
 - ⇒ diagnosis improvement and definition of new specific therapies
- Bio-discovery
 - gene network discovery
 - analysis of multifactorial genetic pathologies
- Pharmacogenesis
 - lab design of new drugs for genic therapies

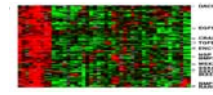
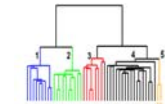



How can data mining contribute?

DBGM 8

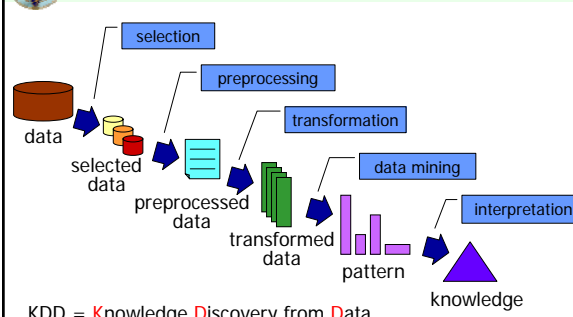
Data mining contributions

- Pathology diagnosis
 - classification
- Selecting genes involved in a specific pathology
 - feature selection
 - clustering
- Grouping genes with similar functional behavior
 - clustering
- Multifactorial pathologies analysis
 - association rules
- Detecting chemical components appropriate for specific therapies
 - classification

DBGM 9

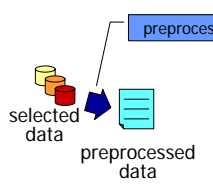
Knowledge Discovery Process



KDD = Knowledge Discovery from Data

DBGM 10

Preprocessing



data cleaning

- reduces the effect of noise
- identifies or removes outliers
- solves inconsistencies

data integration

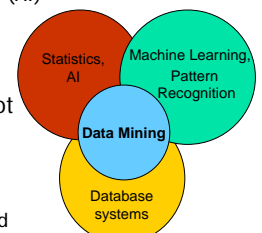
- reconciles data extracted from different sources
- integrates metadata
- identifies and solves data value conflicts
- manages redundancy

Real world data is "dirty"
Without good quality data, no good quality pattern

DBGM 11

Data mining origins

- Draws from
 - statistics, artificial intelligence (AI)
 - pattern recognition, machine learning
 - database systems
- Traditional techniques are not appropriate because of
 - significant data volume
 - large data dimensionality
 - heterogeneous and distributed nature of data



From: P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining"

DBGM 12



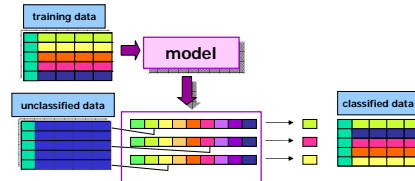
Analysis techniques

- Descriptive methods
 - Extract interpretable models describing data
 - Example: client segmentation
- Predictive methods
 - Exploit some known variables to predict unknown or future values of (other) variables
 - Example: "spam" email detection



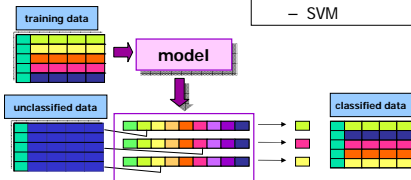
Classification

- Objectives
 - prediction of a class label
 - definition of an interpretable model of a given phenomenon



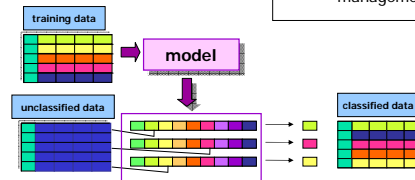
Classification

- Approaches
 - decision trees
 - bayesian classification
 - classification rules
 - neural networks
 - k-nearest neighbours
 - SVM



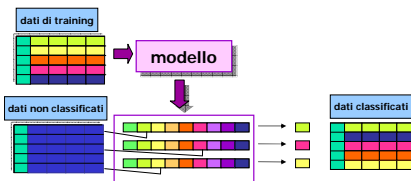
Classification

- Requirements
 - accuracy
 - interpretability
 - scalability
 - noise and outlier management



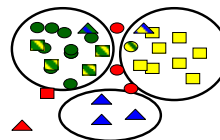
Classification

- Applications
 - detection of customer propension to leave a company (churn or attrition)
 - fraud detection
 - classification of different pathology types
 - ...



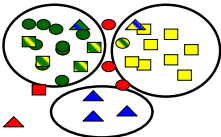
Clustering

- Objectives
 - detecting groups of similar data objects
 - identifying exceptions and outliers



Clustering

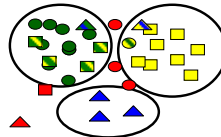
- Approaches
 - partitional (K-means)
 - hierarchical
 - density-based (DBSCAN)
 - SOM
- Requirements
 - scalability
 - management of
 - noise and outliers
 - large dimensionality
 - interpretability



DBGM 19

Clustering

- Applications
 - customer segmentation
 - clustering of documents containing similar information
 - grouping genes with similar expression pattern
 - ...



DBGM 20

Association rules

- Objective
 - extraction of frequent correlations or pattern from a transactional database

Tickets at a supermarket counter

| TID | Items |
|-----|----------------------------|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diapers, Milk |
| 4 | Beer, Bread, Diapers, Milk |
| 5 | Coke, Diapers, Milk |
| ... | ... |

- Association rule
 - diapers \Rightarrow beer
 - 2% of transactions contains both items
 - 30% of transactions containing diapers also contains beer

DBGM 21

Association rules

- Applications
 - market basket analysis
 - cross-selling
 - shop layout or catalogue design

Tickets at a supermarket counter


| TID | Items |
|-----|--------------------------------|
| 1 | Bread, Coca Cola, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coca Cola, Diapers, Milk |
| 4 | Beer, Bread, Diapers, Milk |
| 5 | Coca Cola, Diapers, Milk |
| ... | ... |

- Association rule
 - diapers \Rightarrow beer
 - 2% of transactions contains both items
 - 30% of transactions containing diapers also contains beer

DBGM 22

Other data mining techniques

- Sequence mining
 - ordering criteria on analyzed data are taken into account
 - example: motif detection in proteins
- Time series and geospatial data
 - temporal and spatial information are considered
 - example: sensor network data
- Regression
 - prediction of a continuous value
 - example: prediction of stock quotes
- Outlier detection
 - example: intrusion detection in network traffic analysis



DBGM 23

Open issues

- Scalability to *huge* data volumes
- Data dimensionality
- Complex data structures, heterogeneous data formats
- Data quality
- Privacy preservation
- Streaming data

DBGM 24