

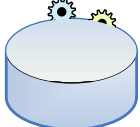

Data mining fundamentals



Elena Baralis
Politecnico di Torino

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*

4

Data analysis





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

2

Example: biological data

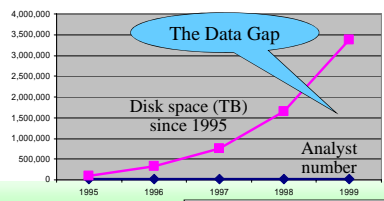
- Microarray
 - expression level of genes in a cellular tissue
 - various types (mRNA, DNA)
- Patient clinical records
 - personal and demographic data
 - exam results
- Textual data in public collections
 - heterogeneous formats, different objectives
 - scientific literature (PubMed)
 - ontologies (Gene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


5

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"




3

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

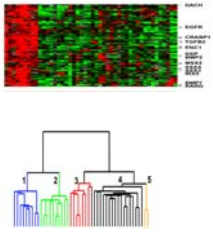
How can data mining contribute?



6

Data mining contributions

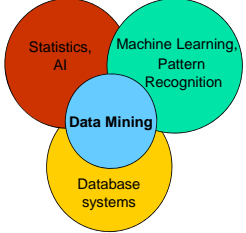
- Pathology diagnosis
 - classification
- Selecting genes involved in a specific pathology
 - feature selection
 - clustering
- Grouping genes with similar functional behavior
 - clustering
- Multifactorial pathologies analysis
 - association rules
- Detecting chemical components appropriate for specific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



DBGM 7

Data mining origins

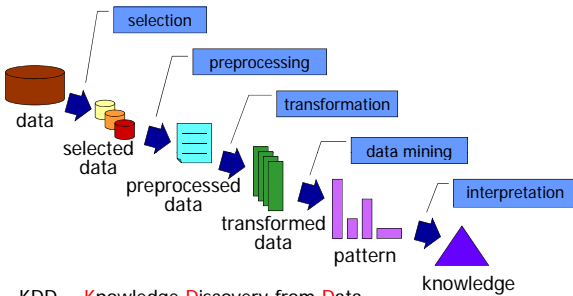
- Draws from
 - statistics, artificial intelligence (AI)
 - pattern recognition, machine learning
 - database systems
- Traditional techniques are not appropriate because of
 - significant data volume
 - large data dimensionality
 - heterogeneous and distributed nature of data



From: P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining" 10

DBGM

Knowledge Discovery Process



KDD = Knowledge Discovery from Data

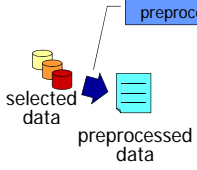
DBGM 8

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

DBGM 11

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

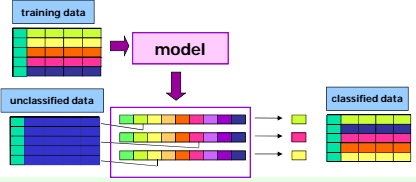
Real world data is "dirty"

Without good quality data, no good quality pattern

DBGM 9

Classification

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



DBGM 12

Classification

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

The diagram shows a flow from 'training data' (a grid of colored squares) to a 'model' (a purple box). An arrow points from the 'model' to 'unclassified data' (another grid of colored squares). From 'unclassified data', arrows point to 'classified data' (a grid where each square is a single color, representing the result of the model's classification).

DBG

13

Clustering

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

The diagram shows several groups of data points enclosed in circles. One group contains green circles, another contains yellow squares, and a third contains blue triangles. There are also several red triangles scattered around, representing outliers.

DBG

16

Classification

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

The diagram shows a flow from 'training data' (a grid of colored squares) to a 'model' (a purple box). An arrow points from the 'model' to 'unclassified data' (another grid of colored squares). From 'unclassified data', arrows point to 'classified data' (a grid where each square is a single color, representing the result of the model's classification).

DBG

14

Clustering

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

The diagram shows several groups of data points enclosed in circles. One group contains green circles, another contains yellow squares, and a third contains blue triangles. There are also several red triangles scattered around, representing outliers.

DBG

17

Classification

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

The diagram shows a flow from 'dati di training' (a grid of colored squares) to a 'modello' (a purple box). An arrow points from the 'modello' to 'dati non classificati' (another grid of colored squares). From 'dati non classificati', arrows point to 'dati classificati' (a grid where each square is a single color, representing the result of the model's classification).

DBG

15

Clustering

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

The diagram shows several groups of data points enclosed in circles. One group contains green circles, another contains yellow squares, and a third contains blue triangles. There are also several red triangles scattered around, representing outliers.

DBG

18


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



19

Open issues

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



22


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



20

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

21