Opening the Black Box: Deriving Rules from Data

Elena Baralis

Politecnico di Torino

Data Base and Data Mining Group of Politecnico di Torino

BICOD 2017 – RuleML+RR 2017

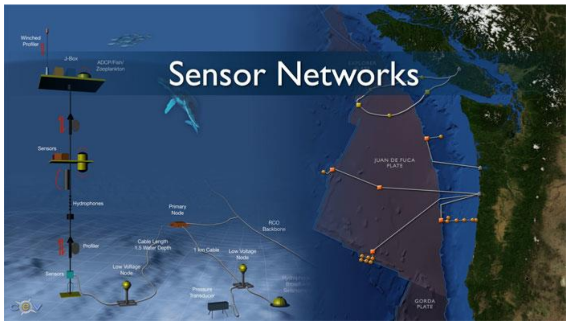
London, July 12th 2017

Opening the black box

i.e., the long and winding road that takes to knowledge…

Mining large datasets

- Continuously growing amounts of data are being collected and stored



Sensor data

User generated content

E-commerce data

- Available for exploration and analysis
  - Patterns and models are extracted from data to
    - describe their characteristics
    - predict variable values

The Quest for Interpretability

- Powerful analysis techniques are being designed
  - Among them, deep learning techniques
- Unfortunately, many high quality models are characterized by being hardly interpretable
  - Data interpretability is important for decision making
- Rules mined from data may provide easily interpretable knowledge
  - both for exploration and classification (or prediction) purposes

Extracting Meaning from Data

- Introducing some types of rules inferred from data
  - association rules
  - (associative) classification rules
  - with variations on the theme…
- Discussing their capability of
  - describing phenomena
  - giving meaning to the data under analysis

Rule patterns

- High quality patterns derived bottom-up
  - Not assuming any apriori knowledge on data
    - will relax somewhat this hypothesis
- Several kinds of pattern
  - descriptive patterns: association rules & itemsets
  - rule models for prediction
- Focus on association rules
- Many application domains
  - Data exploration and explanation
  - Constraint derivation
  - …

Descriptive patterns

- Many different types of association-based patterns
  - itemsets
  - association rules
  - weighted association rules
  - generalized association rules

Association rules

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

Purchases 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 contain both items
  - 30% of transactions containing diapers also contain beer

Transactional formats

- A transaction can be any set of items
  - Market basket data
  - Textual data
    - A document is a transaction
    - Words in a document are items in the transaction

  - Structured data
    - A table row is a transaction
    - Pairs (attribute, value) are items in the transaction
    - Example

Refund=no, MaritalStatus=married, TaxableIncome<80K, Cheat=No

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Married	< 80K	**No**

Example from: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Association rules

- Identification of hidden correlations among data

$$\mathbf{X} \rightarrow \mathbf{Y}$$

- X and Y are itemsets, sets of one or more items
- Quality indices
  - *Support:* percentage of transactions containing X and Y
  - *Confidence:* conditional probability of finding Y given X
  - *Lift:* ratio between rule confidence and support of Y

Considering weight

- Items may be characterized by different importance within a transaction
  - Examples: product quantity, term frequency of occurrence, tf-idf
- Weighted dataset
  - Each item is assigned a weight measuring its relevance in the corresponding transaction
- Weighted itemsets represent correlations among multiple highly relevant terms
  - Several different definitions of weighted itemset support

Document summarization

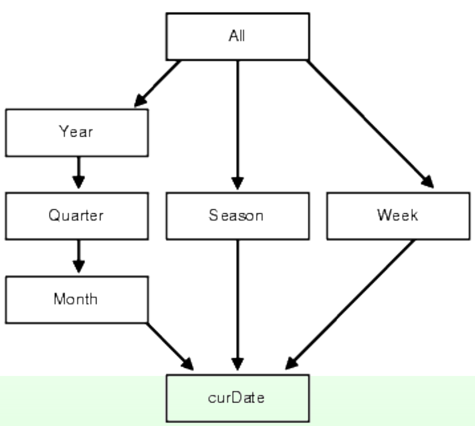
- The summary of a collection of news documents ranging over the same topic
  - provides a synthetic overview of the most relevant news facets
  - does not require access to the entire document collection
- Itemset-based summarizers analyze the co-occurrences between multiple document terms
  - frequent weighted itemsets consider only the correlations between *highly relevant* terms
  - term weights measure term relevance in the analyzed collection

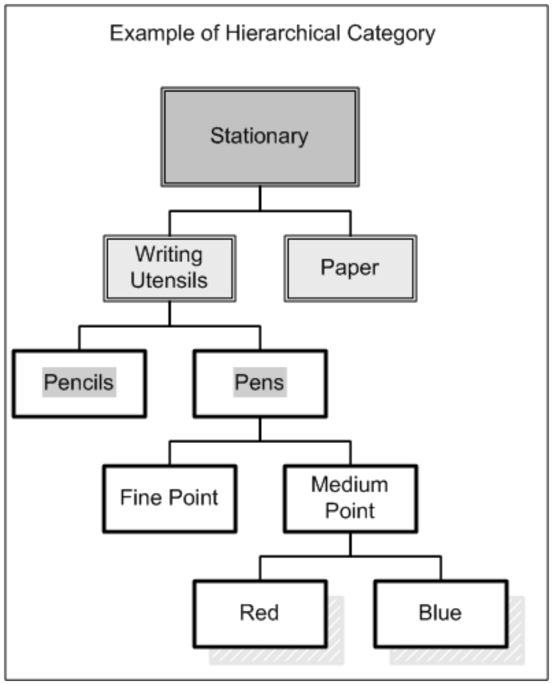
Document summarization

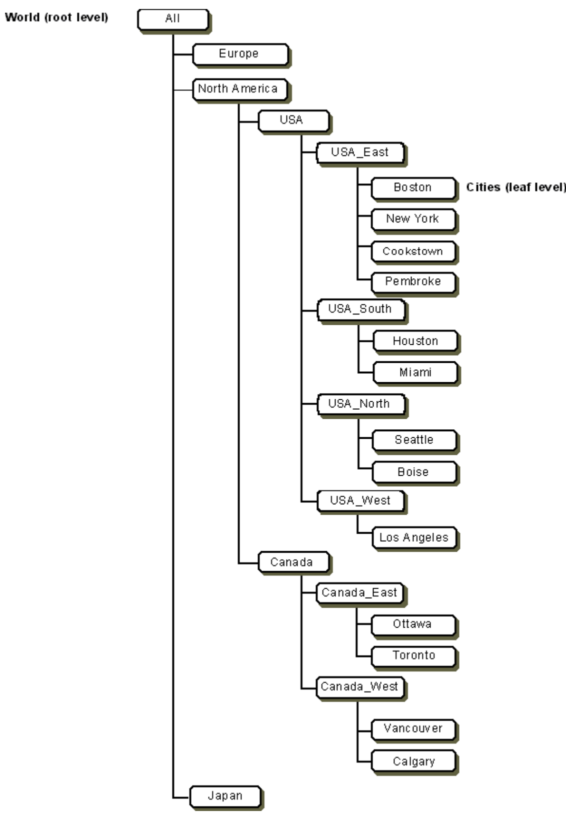
- Language-agnostic approach
  - makes minimal use of language-dependent analyses (stopwords, optionally stemming)
  - is easily applicable to document collections written in different languages (Arabic, Czech, English, French, Greek, Hindi)
- Item weights are particularly effective for summarizing documents written in languages other than English

Considering hierarchies

- Generalization hierarchies
  - Aggregation over attributes in a dataset
  - Typically user provided
- Examples
  - Time hierarchy
  - Product category
  - Location hierarchy
  - …

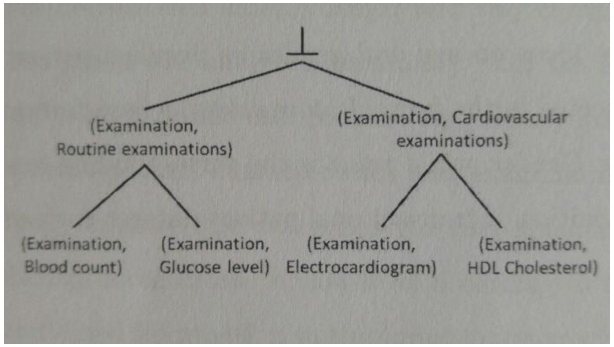






Taxonomy

- A taxonomy is a set of is-a hierarchies that aggregate data items into higher-level concepts
- Data item
  - Instance in the (transactional) dataset
  - Represents detailed concepts
- Generalized item
  - Aggregation in higher-level concepts
  - Represents abstractions on instances



Generalized itemsets

- Sets of items at different generalization levels
  - May contain data items together with generalized items defined in the taxonomy
  - Summarize knowledge represented by a set of lower-level descendants
    - Both frequent and infrequent
- A generalized itemset covers a transaction when all
  - its generalized items are ancestors of items included in the transaction
  - its data items are included in the transaction
- Generalized itemset support
  - ratio between number of covered transactions and dataset cardinality

Context-aware data analysis

- Context data provided by different, possibly heterogeneous, sources
  - Mobile devices provide information on
    - the user context (e.g., GPS coordinates)
    - the supplied services
      - temporal information
      - service description
      - duration
  - Additional information available
    - demographics of the user requesting the service

Generalized itemset example

user: **John,** ***time:*** **6.05 p.m.,** ***service:*** **Weather**
(s = 0.005%)

- A very low support
  - The itemset may be discarded
- By generalizing
  - the time attribute on a time period
  - the user on a user category

user: **employee,** ***time:*** **6 p.m. to 7 p.m.,** ***service:*** **Weather**
(s = 0.2%)

- May discover interesting properties generalizing *infrequent* items

Generalized association rules

- Extension of "classical" association rules

$$X \rightarrow Y$$

- X and Y are either generalized or not generalized itemsets
  - Support, confidence and lift are defined accordingly

Patient data analysis

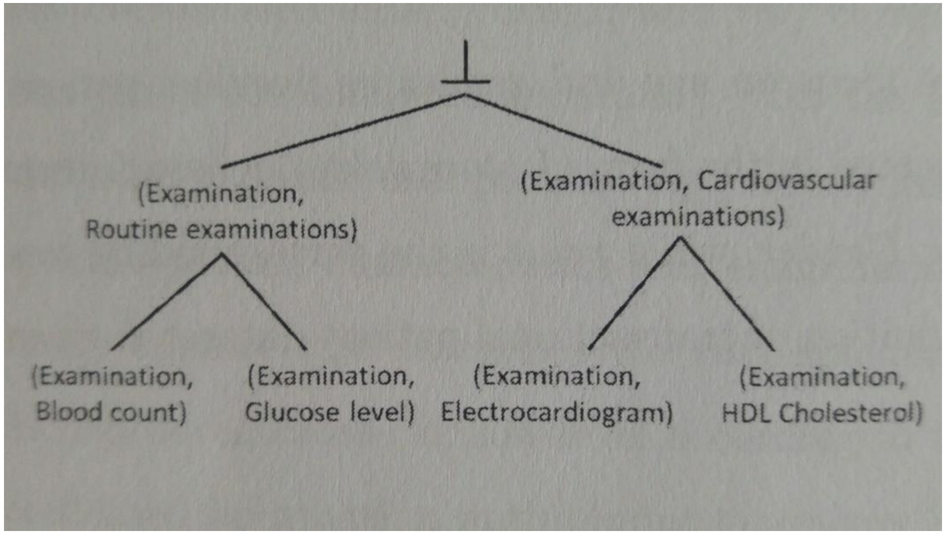
- Analysis of multiple level correlations on patient treatment historical data
  - Dataset collected by an Italian Local Health Center
    - Diabetes complications at various severity levels
    - 95K records, 3.5K patients
  - Features
    - Prescribed examinations (26 examinations, 7 categories)
    - Prescribed drugs (200 drugs, 14 categories)
    - Census patient data (gender, age discretized in age groups)
- Sparse dataset
  - Difficult setting of support threshold
    - Low: generates too many rules
    - High: interesting information at lower levels of abstraction may remain hidden

Patient data analysis

- Rule exploration in top-down fashion
  - From small subset of high-level rules drill down to more specific rules
    - Descending level of abstraction on the considered taxonomy
  - Discovery of rule groups at different abstraction levels
    - Typically more manageable for manual exploration
- Consider only non redundant rules
  - Compact subset based on closed itemsets
    - Rule is redundant if it has same support and confidence of its specialized version
  - Reduces cardinality of rule set

Patient data analysis

- High-level rules
  - Only generalized itemsets (examination and drug categories)
  - Represent general knowledge
    - May be too high level to perform targeted analyses

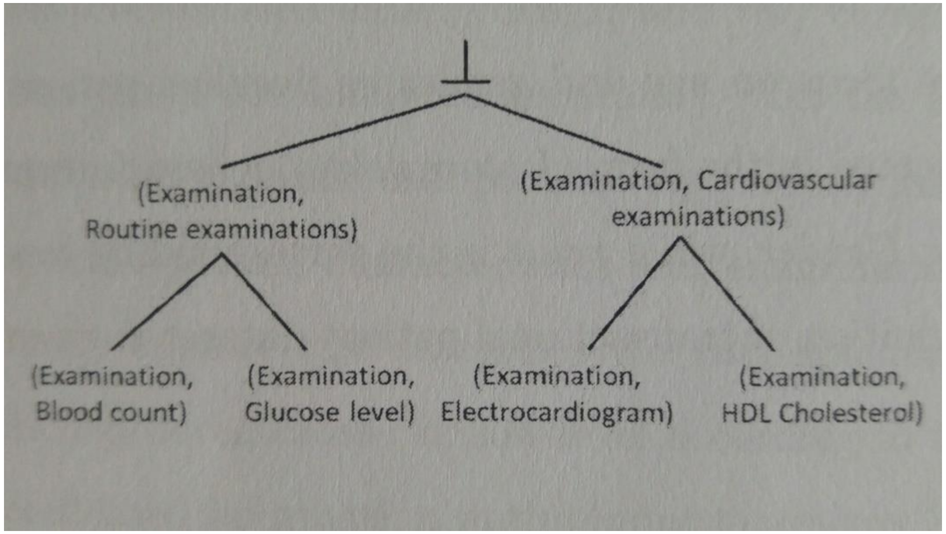


Patient data analysis

- Extracted high-level rule

  (Examination, Liver) -> (Examination, Kidney)

  - Frequently prescribed together
  - May be used for examination scheduling



Patient data analysis

- Cross-level rules
  - Different abstraction levels (generalized items and data items)
  - Combine detailed and general information
- Extracted cross-level rule

  (Examination, Liver) -> (Examination, Uric acid)

  - Insight into specific kidney examinations correlated with liver examinations
    - Confidence: 74.8%

Patient data analysis

- Low-level rules
  - Only not generalized itemsets (only data items)
  - Very detailed knowledge
    - Covered by high and cross-level rules
  - Large rule set
    - Challenging exploration task
  - Drill down exploration based on formerly extracted high and cross-level rules

Outcomes

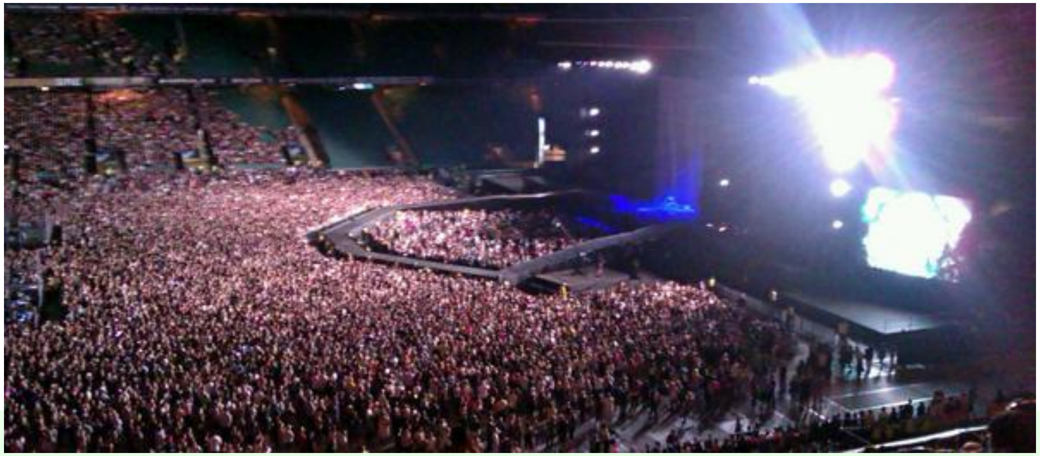
- Allow experts to
  - Identify medical treatments commonly followed by patients with a given disease
  - Verify adherence of medical treatment to shared medical guidelines
  - Improve the effectiveness of medical treatments
  - Plan resource allocation and reduce costs incurred by organization

Flipping correlations

- Discovery of contrasting situations between ancestor and descendant itemsets
  - Identify exceptional or unexpected situations
- Itemsets characterized by a correlation type
  - Positive, negative, or null
  - Correlation strength measured by correlation indices
    - Kulczynsky, lift, …
- Itemsets whose *correlation type flips* (changes) when its items are generalized to a higher level of abstraction

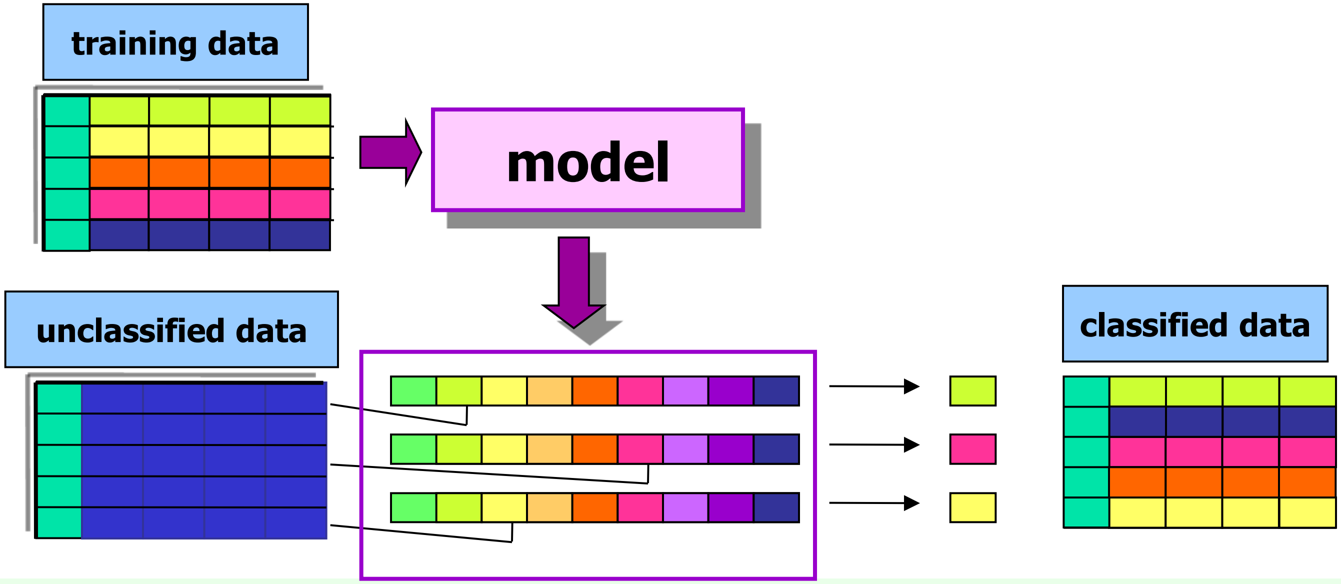
Flipping correlations

- Twitter dataset on Music topic
- Flipping correlation
  - Generalized itemset, *negative* correlation
    (Date: Working day), (Location: Twickenham Rugby Stadium)
  - Exception, *positive* correlation
    (Date: 2012-09-08), (Location: 51.45542-0.34165)
  - Lady Gaga sold-out concert in the stadium on 2012-09-08



Classification

- Objectives
  - prediction of a (discrete) class label
  - definition of a model of a given phenomenon
    - interpretable?



Rule patterns for classification

- Targets
  - Defining an *interpretable* rule *model* capable of
    - assigning class label to unclassified object
    - describing main class characteristics
  - Providing *reasons for* a *classification* outcome
    - why a class label is assigned to a given instance?
- Several types of rules
  - Class association rules (CARs)
    - Good quality classification models
    - Many different approaches

Class Association Rules

- Structure of "classical" association rules

$$X \rightarrow Y$$

  where Y is a class label
- CARs selection
  - Rule selection & sorting
    - based on support, confidence and lift thresholds
  - Rule pruning
    - Database coverage: the training set is covered by selecting topmost rules according to previous sort

Lazy pruning

- L^3: Live and Let Live
  - Low support threshold for rule extraction
    - Rule selection based on confidence
  - Multiple support thresholds for different classes
  - Level-based approach in selecting rules
    - *Good rules,* small subset of high quality rules
    - *Spare rules,* larger set of rules not used during database coverage
    - *Harmful rules,* discarded because only wrongly classifying training data
- High quality model
  - Larger rule set, considering spare rules

Instance-centric approaches

- DeEPs
  - Emerging Pattern are patterns that sharply differentiate one (training) class from the others
    - Interesting patterns occur frequently in one class and less frequently in the others
  - *Lazy* classification
    - EP extraction takes place for the given test instance
    - Aggregate supports of extracted EPs assign class label
- Harmony
  - Selects a subset of best possible rules for each training instance
    - highest confidence frequent covering rules

The challenge of big data

- Huge data collections exacerbate the problem
  - Very sparse datasets
    - Support threshold setting
  - Computational challenge
    - Scalability in item cardinality is a challenge
    - Hadoop/Spark framework not straightforwardly usable
- Local exploration of datasets
  - Several criteria to select area to explore
    - Rule constraints
      - Schema constraints, item constraints
    - Predicates on attributes
    - Instance constraints

Rules as building blocks

- Rules may support black box learning paradigms
  - Learn rule patterns from data
    - learn some abstractions by experience
    - e.g., a positioning rule for objects
  - Use learned patterns (abstractions) to support deep learning techniques
    - drive learning also by abstractions
    - e.g., use rule to improve object detection
- It is the way our brain works!

Thanks to…

- everybody in my research group, PhD students and researchers, but especially…
  - Giulia Bruno
  - Luca Cagliero
  - Tania Cerquitelli
  - Silvia Chiusano
  - Paolo Garza
  - Luigi Grimaudo

Thank you!

Questions?