Bioinformatics

Profiles data mining

Martin Saturka

http://www.bioplexity.org/lectures/

EBI version 0.4

Intriguing on profiles

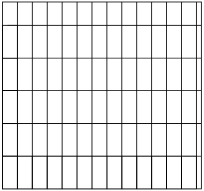
Data mining technics for relation exploration on profiles.

- associations, multitudinal quantifiers, dinorms.
- bootstrapping, permutation tests. entropy.

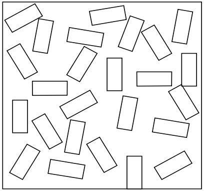
Main topics

- mining technics
  - data and knowledge
  - important information
- observational calculi
  - fuzzy logic, quantifiers
  - aggregation functors
- multivariate statistics
  - resampling methods
  - inference, decisions

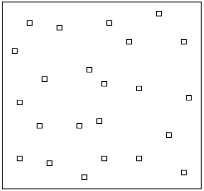
Relation types



Solid state

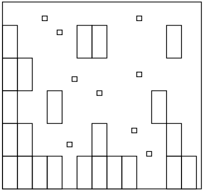


Fluid state

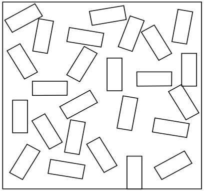


Gas state

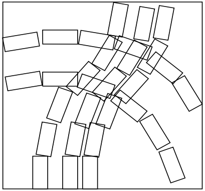
search for the 'organismal' relations:



A form



B form



C form

Data mining targets

- general features
    - great fraction of objects contains the property
    - common properties, object comprehension
        - frequent higher / lower expressions of some genes
- unusual events
    - rarely occuring 'alarm-trigger' situations
    - regular checking, after comprehension is done
        - not to allow cellular behavior to go out of frames
- specific features
    - fraction difference between object groups
    - partial object class characterization
        - tracking expression of many genes → moderate between-group expression differences acceptable

Knowledge expression

formulation of knowledge about explored systems

- statements, recommendations
  - what something either is or is supposed to be
- probabilities, beliefs, degrees of truth
  - various typs of uncertainty expression
- data tables, databases, protocols
  - actual data / information / knowledge storage

Hypotheses

hypothesis creation and testing

- unkown relations
  - when we do not know what to expect from the data
  - search for every important feature and property
  - hypothesis creation processed by data-mining technics
- supposed relations
  - when we have an alleged factual hypothesis
  - estimation of particular statement plausibility
  - hypothesis testing processed by statistics technics

Fourfold tables

	ψ	$\neg\psi$
φ	*a*	*b*
$\neg\varphi$	*c*	*d*

- event table rationality
  - counts of event cases: *a*, *b*, *c*, *d*
  - associations between (φ, ψ) data features φ, ψ for e.g. particular gene expressions
  - non-informational data: usually *c*, *d*
    - the case of ’nothing happens’ situations
    - expressions of ’for many φ having many ψ’

Formulae

logic basics

- predicate and observational calculi
  - formulae $\varphi(x)$, $\psi(x)$, $\varphi(x) \wedge \psi(x)$, $Q(\varphi, \psi)$
  - quantifiers $Q(\varphi, \psi)$
    - φ antecedent, ψ succedent - consequent
  - variables: x for particular experiments, tissues
    - supposed implicitely if not written

- expression of important pieces of information
  - directional associations
    'for many φ having many ψ'
  - mutual associations
    'few situations of single φ or ψ'

Crisp and fuzzy data

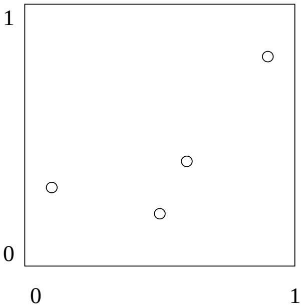
various meanings of fuzziness

usually when something is not defined as binary 0/1 values

- mathematical logic
  - precise mathematical meaning of fuzziness
  - formula evaluations in the whole $[0, 1]$ interval
  - specific axioms for particular fuzzy logics
- connectives, t-norms
  - *and* connective defined by (continuous) t-norms, i.e. $t(x, y)$ mappings $[0, 1] \times [0, 1] \rightarrow [0, 1]$ that are commutative, associative, non-decreasing, $t(1, y) = y$
  - implications - residual: $i(x, y) = z$ mappings for $\max\{z \mid t(x, z) \leq y\}$

Fuzzy events

fuzzy data	A	B
1	0.1	0.3
2	0.9	0.8
3	0.6	0.4
4	0.5	0.2



- event squares as the fuzzy-data analogy of fourfold tables

Multitudinality

implication-like quantification reduced on important data

- more vs. less important events
  - the succedent under the antecedent condition
  - not ot be overwhelmed by nothing-happens data
    - feature pairs where is valid: if the antecedent is satisfied than succedent is usually satisfied too
    - if the antecedent is not satisfied than we do not care
- crisp data-case multitudinality
  - quantifiers are defined with the help of fourfold table values a, b, c, d (a is the count of event where both antecedent and succedent are satisfied, b is for just the antecedent satisfied, etc.)
  - if Q is satisfied on (a_1, b_1, c_1, d_1) data and we have $a_1 \leq a_2$, $b_1 \geq b_2$ for another data (a_2, b_2, c_2, d_2) than Q is satisifed on the 2-indexed data as well.

Fuzzy multitudinality

directional multitudinal quantifiers Q on fuzzy data

- the events are not 'yes'/'no' situations
  - generalizing the definition for 'something partially happens' event cases

- Q definition with the help of event squares
  - $\{x, y_1\} \rightarrow \{x, y_2\}$ for $y_1 \leq y_2$ does not decrese Q value
  - addition / removal of $\{0, y\}$ events does not change Q value
  - $\{1, 1\}$ is the best event for the Q valuation
  - $\{1, 0\}$ is the worst event for the Q valuation

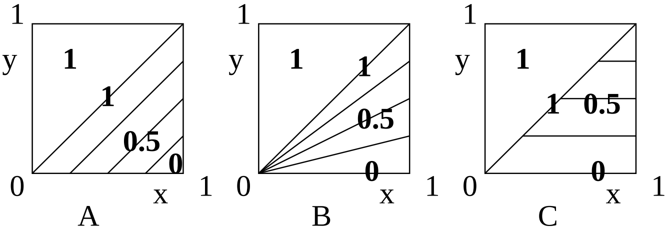
Quantifiers

- multitudinal quantifiers based on residual implications
  - product t-norm: $t(x, y) = x \cdot y$, $i(x, y) = 1$ for $y \geq x$, or y/x
  - Lukasiewicz t-norm: $t(x, y) = max(0, x + y - 1)$, $i(x, y) = 1$ for $y \geq x$, or $1 + y - x$
  - Goedel t-norm: $t(x, y) = min(x, y)$, $i(x, y) = 1$ for $y \geq x$, or y

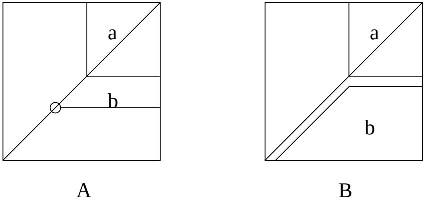
- particular multitudinal quantifiers
  - weighted implication means
    - the product t-norm case $\sum \min\{x, y\} / \sum x$
  - weighted implication quantiles
    - each event has its length according to the x value
  - quantile estimations
    - analogically to the standard quantile estimation
  - survival models
    - modified version of the Kaplan-Meier estimator

Examples

implication functions



survival based quantification



Symmetric notions

- mutual multitudinal quantifiers
  - up to know, we had directional quantifiers (i.e. relations)
  - directionality to bidirectionality switch by taking both direcions into account
  - taking the lesser values during weighted implication computations for $\{x, y\}$ and $\{y, x\}$ events
- distances
  - symmetric multitudinality can be used as a feature to feature (e.g. inter-genes) similarity measure
  - the product t-norm case: $\sum min\{x, y\} / \sum max\{x, y\}$

Clustering

relations available - use them

- too many gene formulas (genes, gene tuples)
  - cluster the gene formulas into groups of similar expressions
  - mutual multitudinal quantifiers available as a similarity measure

- directional clustering
  - classical centered clustering
    the items as both greater and lesser than respective centers
  - uni-directionality centering
    one set of lesser and one set of greater items per cluster

Gene selection

which gene combinations are substantial for tissue distinction

- (formula based) class covering
  - selection of gene formulas which are highly valuated at an object class
  - go through subsequently longer formulas φ_1, $\varphi_1 \wedge \varphi_2$, $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$, ... while a class has a high valuation and no other class has a lower valuation
  - set of many such reached formulas forms a pool of distinction gene expression properties

- gene shaving (based on PCA)
  - 'shaving off' genes with low dot product to an eigenvector
    - the rest genes used for PCA recomputation iteration
  - the right group size by the gap statistics
    - the most variance explained - compared to random groups

Feature aggregation

putting many features into single property

- aggregators
  - trend prediction from many symptoms
  - demands: continuity, stability, associativity
  - one way trand by e.g. logical connectives

- directions
  - single or two opposite proneness directions
  - usually some combination of antagonistic trends
  - stable continuity with associativity impossible

Regulatory calculus

gene regulation description

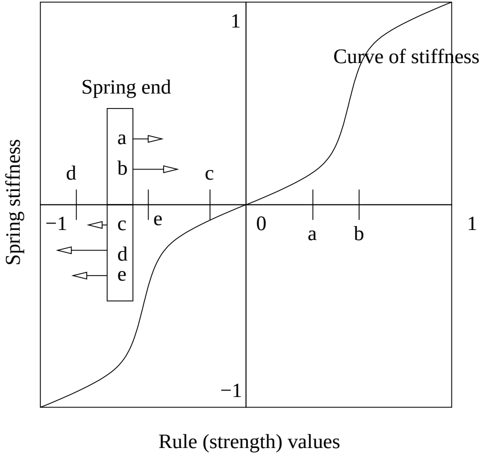
- value separation
  - both gene activation and inhibition are important events
  - making statement pairs out of single gene expression statements
    - old: gene$_i$ expression
    - new: gene$_i$ activation, gene$_i$ inhibition
- description formulas
  - addition of the 'opposite' connective to the language
  - example: $g_1 \wedge g_2 \wedge$ oppg_3
    i.e. both genes 1 and 2 are activated and gene 3 is inhibited

Dinorms

alike trends first, then overall combination

- trend aggregation
  - aggregations made separately for the opposite directions
  - finally, combination of the overall anti-directional trands
  - stability, continuity and a kind of associativity gained
- symbolic notation
  - $(A_{upp} \hookleftarrow A_{low}) \vee (A_{low} \hookleftarrow A_{upp})$
  - the overall combination by coimplications
    - kind of difference measurement
    - one of the two items is zero, the other one is the overall trend

Dinorm structure



- schematic spring example of a general dinorm

Resampling methods

amounts of data features and relations

- based on empirical sampling distribution
  - make an assumption of samples interchangebility
- resampling technics
  - bootstrapping
    - trend evaluation under multiple symptoms
    - leading into a real valued parameter estimation
  - permutation tests
    - p-value estimation under unknown data distribution
    - difference between two groups estimation
  - gap statistics
    - (cluster) size choice from a sequence of ranks
    - point of the largest group-plausibility

Resampling algorithms

- bootstrap method
  - take all the data samples as an unordered set
  - make a new sampling - with replacements - of the original data size
  - compute the explored property as usually, save the value
  - make the new sampling / computation many times
  - the new distribution is the one of the property,
    it tends to be the normal distribution

- permutation tests
  - take all the data samples as an unordered set
  - separate the data by accident into right-sized groups
  - compute the explored difference, save the value
  - make the new sampling / computation for many times
  - the new distribution is the one of the test
  - p-value as the ratio of larger gained differences

Multiple decisions

- parallel subdecisions
  - nearest neighbors
- sequential subdecisions
  - classification trees

- after data mining is done
  - new features gained for decision making
  - nearest neighbor search should be improved
  - features available for multiple decisioning

- missing data
  - some experimental data missing in virtually every dataset
  - lost subdecisions either ignored or modelled by the most similar samples

Trees

questions → answers → classification

- C&RT
  - classification and regression trees
  - tree: each (non-leaf) node is an if-then-else condition
  - subsequent questions / answers to make classification
  - leaf nodes are the classes (diagnoses)

- algorithm
  - build tree
    - subsequent best-separation splitting
  - tree pruning
    - to avoid overfit done by too specific tree learning

Forests

random forests approaches

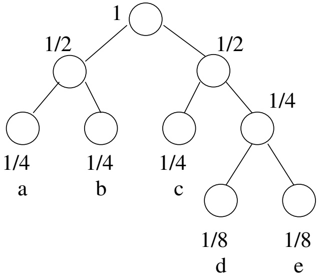
- voting - for multiple classifications
  - each tree has a vote to make a classification
  - outputs of many trees → averaged results
  - could be used as a self-contained data-mining technics

- bagging - bootstrap based aggregating
  - training data being taken by bootstrapping
  - constructing multiple classification trees
  - final result as an averaging consensus

Entropy

probabilistic structure characterization

- Fisher information
  - suitability of an experimental / statistical schema
  - depends on distance measures, keeping track of locality
- Shannon entropy
  - global measure of overall stochasticity
  - length of the most parsimonious alphabet



Splitting

suitable data separations

- subsequent separations
  - classification trees
  - ILP programming
  - clustering trees
- tree structure
  - to make the classification tree resembling the entropy tree
  - not to construct trees too deep
    - many questions → many errors

ILP

relational data-mining on multiple tables

- inductive logic programming
  - data and classes given by positive and negative examples
  - data classes characterization by logic program (hypothesis)
  - for (deterministic) data with limited amount of attributes
- ILP technics
  - to construct hypothesis for
    - the positive examples being proved by it
    - the negative examples not being proved by it
  - separation tree construction
  - prooving by traversal the tree

Learning

- learning **validation**
  - over/under-fitting
    - stop the learning process when the classification success growth changes to be slow / shallow
  - cross-testing
    - to test the learned classification method on independent data
    - split initial data into two groups: for learning and for testing
    - enough data - 1/3 for testing, otherwise 1/10 for testing
- Occam's razor
  - statistical character of the principle
  - do not use more parameters than necessary

Kolmogorov complexity

the absolutely shortest description

theoretical notion with nice pseudoparadoxes

- data compression
  - entropy encoding the first step (for particular symbols)
  - standard compression technics possible if nothing better
  - nearly everything contains some regularities
- MDL
  - minimal description length approach
  - whether to use plain sequence or a found regularity
  - if description of the regularity together with the reduced sequence longer than the old sequence, do not use the weak regularity

Items to remember

Nota bene:

fuzzy logic, crisp and fuzzy data

- Data mining
  - event tables, squares
  - multitudinal quantifiers
- Feature combination
  - aggregators, resampling
  - decisions, entropy