Data Mining in NeuroInformatics

Arno Siebes
Department of Computer Science
Universiteit Utrecht

Data Mining

Inducing Models and Patterns from Databases

Model: a succinct description of the complete database

Pattern: a *local/partial* model

Patterns

- Frequent Patterns, e.g.,
  - frequent item sets (e.g., sets of items people often buy together)
  - frequent sequences (e.g., people tend to rent LOFTR I, II, III in that order)
- Association Rules

$$\textbf{Bread} \ \wedge \ \textbf{Cheese} \ \rightarrow \ \textbf{Milk} \ \ (25\%, 58\%)$$

- Subgroup Discovery

$$\textbf{Age} \ \in [18, 25] \wedge \ \textbf{Sex} \ = male \rightarrow \ P(acc) = 34\%$$

Examples of Patterns

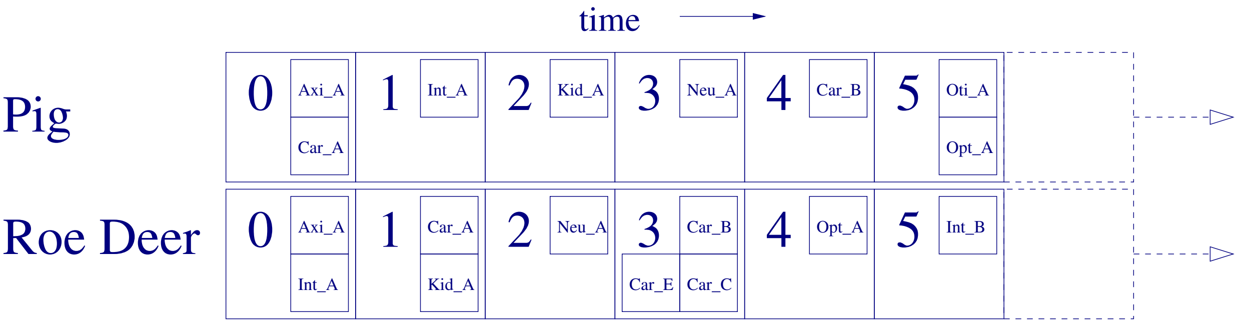
- Areas in the brains that are often active together.
- Activation sequences, both low and high level
- If these two areas are active, this third area will probably also be active
- If these features are present, there is a high probability that the patient has Alzheimer

Developmental Biology

Question:

Is there a relation between species development and evolution?

Data: sequences of events during development:



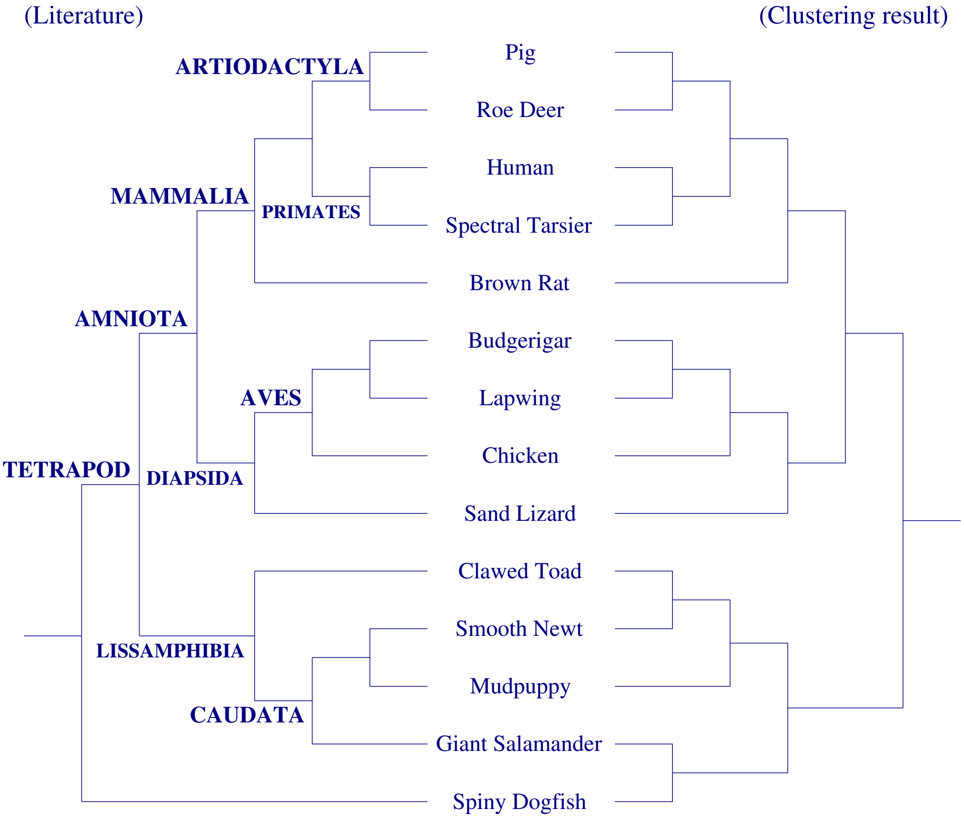
time →

	0	1	2	3	4	5
Pig	Axi_A Car_A	Int_A	Kid_A	Neu_A	Car_B	Oti_A Opt_A
Roe Deer	Axi_A Int_A	Car_A Kid_A	Neu_A	Car_B Car_E Car_C	Opt_A	Int_B

Method

- Compute the frequent sequences
- Use these frequent sequences as features of the species
- Use the Jacquard similarity measure on these featues
- Cluster the species based on this measure

Result


(Literature)
(Clustering result)
ARTIODACTYLA
MAMMALIA
PRIMATES
AMNIOTA
AVES
TETRAPOD
DIAPSIDA
LISSAMPHIBIA
CAUDATA
Pig
Roe Deer
Human
Spectral Tarsier
Brown Rat
Budgerigar
Lapwing
Chicken
Sand Lizard
Clawed Toad
Smooth Newt
Mudpuppy
Giant Salamander
Spiny Dogfish

Models

- Supervised, e.g.,
  - Classification: Trees, Support Vector Machines, ...
  - Regression: Trees, Support Vector Machines, ...
- Unsupervises, e.g.,
  - Clustering: Density Estimation, k-NN, Hierarchical, ...
  - Graphical Models: Bayesian Networks, Conditional Random Fields, ...

Scaling

- A special attention point in data mining is scalability
  - in the number of attributes/features/variables (possibly tens of thousands)
  - the number of tuples/rows/cases (possibly hundreds of millions)
- Even drawing a random sample from a very large database is far from easy...

Only One Table?

- Most data analysis is defined on one input table.
- However, in reality the data comes from multiple tables in multiple databases.
- Moreover, often these tables cannot be integrated without a loss of information.
- For the simple reason that there should be a 1-1 correspondence between rows in the table and the (real world) *objects* we are analysing

Banking Example

The Data: Personal data and Account data

Integration: some possibilties

- join: one person multiple rows
- Each account as a seperate attribute

  1. does it matter which amount at which bank?
  2. exponential blow-up of the table

- What if an account is shared?

In Neural Sciences

Studying Alzheimer

- Various numbers of scans
- Patient data
- Various numbers of diagnostic tests
- Treatment data
- Perhaps microarray data

This never fits one table!

A Common Problem

This is not unique to NeuroInformatics:

- most commercial applications (like the banking example)
- other scientific domains, e.g.,
  - Life Sciences (bioinformatics)
  - Astronomy

The solution is called Relational *mining*

Relational Data Mining

The Data: one or two assumptions

1. the data resides in multiple related tables (keys and foreign keys).
2. for predictive modelling: the *target* table has one row for each object

The Algorithms: "decide" which is the best way to integrate the tables:

- hence, the integration is part of the search
- currently often based on aggregates

State-of-the-Art: some nice algorithms, still lots to do

Distribution

- A major reason for the neuroinformatics initiatives is to share data.
- Is it reasonable to assume that one can ship all data to one site?
- Possible not, because, e.g.,
  - privacy regulations may forbid this
  - the total size would be far too large
- Can we distribute the data mining?

Distributed Data Mining

No Data Shipping Necessary in, e.g.,

- Frequent patterns: a pattern can only be frequent globally if its is frequent in one of the sites
- Bayesian Networks

Data Shipping Necessary: subgroup mining

Research Questions (just two examples)

1. What is the least amount of information to be shipped to allow distributed mining?
2. Can we guarantee privacy protection?

The Volume of Data

- Like in bioinformatics, the potential volumes of data to be analysed in neuroinformatics are staggering.
- This means you will need “big iron”
- Together with data distribution: **The Grid**
- Grid mining is still in its infancy

The Features Revisited

- We have been talking about the features as if it is clear what the good features are.
- However, raw scans are probably not often the right data to mine (don't expect miracles)
- Examples such as well-known brain areas might be much better
- Which features to use is domain knowledge (i.e., I have no clue)
- Mining might suggest new features

Conclusion

NeuroInformatics is an interesting field for data miners

- What features, what patterns, what models?
- Relational mining
- Distributed
  - while preserving privacy
  - large volumes: Grid
- Plenty of room for the development of new algorithms (yummy!)