A Brief Introduction to Data Mining

L. Torgo

ltorgo@dcc.fc.up.pt

Departamento de Ciência de Computadores
Faculdade de Ciências / Universidade do Porto

Sept, 2014

Motivation for Data Mining?

- Data aquisition methods have evolved very rapidly
- Databases have grown exponentially
- These data contain useful information for the organisations
- Size makes manual inspection almost impossible
- Automatic data analysis methods are required to optimise the use of these huge data sets

What is Data Mining?

A possible definition:

> *Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner*
>
> *in* Principles of Data Mining (Hand et.al. 2001)

Searching for relationships

The process of searching for unknown relationships on the data usually involves several steps, like:

- determining the representation of the problem to use
- deciding how to quantify and evaluate/compare how well different representations of the relationships (models) fit the available data
- deciding which data management actions are required to implement the necessary algorithms efficiently

An example

- Suppose we want to understand how the entrance grade of students at the university influences the number of years they take to finish their degree.
- We have collected a data set where each entry contains the grade of a student (a real number) and the number of years it took her(him) to finish the degree (an integer).
- We could decide to fit a linear regression model to the data set - a model of the form $NrYears = \alpha + \beta \times Grade$.
- The degree of fit of this type of models can be calculated by comparing the values predicted by the model against the collected values, and calculating some form of average prediction error
- As the computations required to obtain this type of models are rather simple, most probably no special data management actions would be necessary even for very large data sets.

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Data Mining and Knowledge Discovery in Databases

Data mining is sometimes taken as one of the steps of the process of knowledge discovery

Fonte: Introduction to Data Mining, Tan et.al. (2006)

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Data Sets

- A data set is a collection of measurements taken from some environment.
- In the simplest case, we have p measurements for a set of n objects, i.e. a data matrix of dimension $n \times p$. The n rows represent the objects for which we have collected data. The p columns represent the measurements that were made for each object.
- The rows of the data matrix are also often named examples, instances, records or cases, while the columns are sometimes referred to as variables, features, fields or attributes.

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

An example of a data matrix

Age	Sex	Area	Income
45	m	insurance	85000
32	f	education	72500
24	f	services	97000
.	. . .	. . .	. . .

Table : An example of a data table (matrix)

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Types of Measurements

- Quantitative measurements
  - Integer values
  - Real numbers
- Categorical measurements
  - Ordinal variables (implicit ordering among values - small, medium, large)
  - Nominal variables (no order - red, blue, yellow)

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Types of Data Sets

- Simple data tables (the most common situation)
- Databases (multiple data tables related with each other)
- Data streams, time series
- Text
- Multimedia data (images, sound, etc.)
- etc.

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Type and structure of the models

- Global
- Local

- Mathematical formulae
- Logical formulae
- Black boxes
- etc.

Different models frequently lead to different compromises in terms of understandability and predictive accuracy

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Examples of different models

- Logical formulae - decision rules

```
IF amount = high AND salary = low AND employment = short.term
THEN risk = high

IF amount = average AND salary = high
THEN risk = low
```

- Mathematical formulae

```
houseValue = 10.5 + 5.2 * nrRooms - 3.1 * distCenter + 2.6 * area
```

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Some of the Main Data Mining Tasks

- Exploratory Data Analysis
  - summarisation and visualisation tools
- Descriptive Models
  - probabilistic models
  - clustering models
  - association rules
  - anomaly and deviation detection
- Predictive Models
  - classification models
  - regression models

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

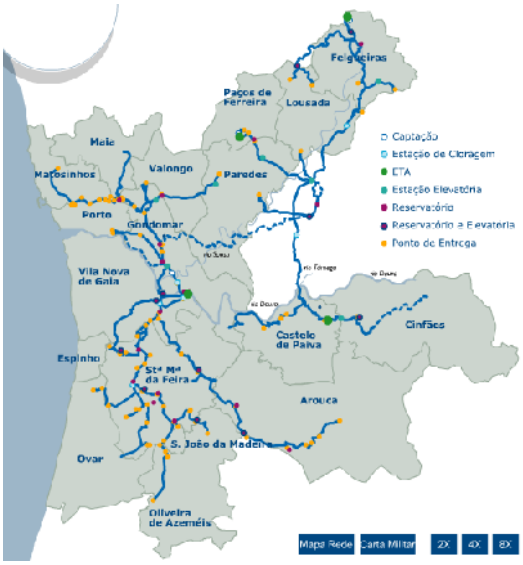
The Key Issues on a Data Mining Project

- Data Structure
  - what to measure? pre-processing steps? ...
- Model Structure
  - what type of model(s) should we build? ...
- Score Function
  - how to evaluate the obtained models? ...
- Optimisation and Search Method
  - how to search and optimise the models in the context of the selected structure? ...
- Data Management Strategy
  - how to handle the data efficiently during model construction/evaluation? ...

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

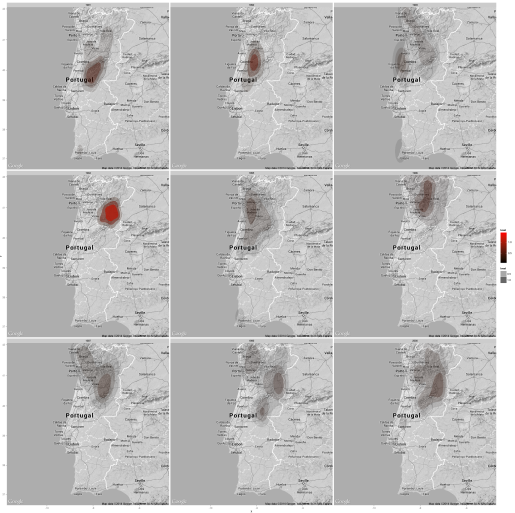
Monitoring and Forecasting Water Quality Parameters

- Hundreds of water quality parameters
- Legal limits to obey
- Heavy fines associated with limits
- Strong socio-economical impact
- Two main data mining tasks:
  - Monitoring what is happening
  - Predicting future events



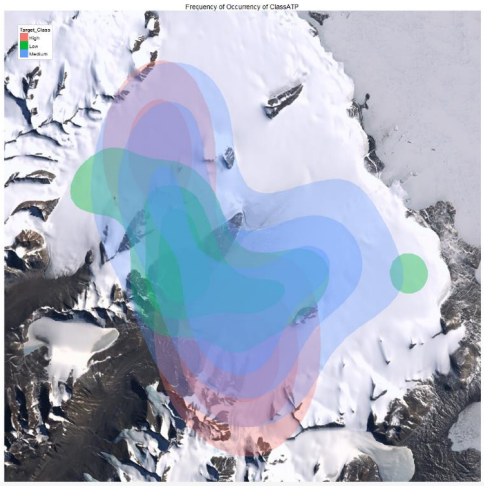
Monitoring and Forecasting Forest Fires

- Problem with a strong socio-economical impact
- Identify the key drivers for fire occurrence
  - Socio-demographic factors
  - Landscape characteristics
  - Meteorological factors
  - etc.
- Identify trends
- Forecast problems



Data from an International Expedition to Antarctica

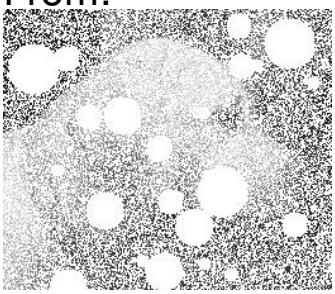
- Identify key factors for the existence of life under extreme conditions
  - Identification of geographical regions with high relevance in terms of bio-diversity



Spatial Interpolation Methods

- Forecast values of variables at places where no data is available
- Sampling cost reduction
- Potential applications:
  - Security
  - Biology
  - etc.

From:



To:

Fraud Detection

- Fraud detection usually involves auditing activities
  - These have costs and are resource-bounded
- Detected frauds have different outcomes (results/benefits)
- How to apply the limited auditing resources to the most promising cases?

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Monitoring and Forecasting Machine Failures

- Industrial machines with several attached sensors measuring different parameters
- Different production contexts
  - What is *normal* varies with the context
- How to anticipate machine failures to take preventive actions?

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO