Predicting Algae Blooms

- brief description of this case study -

L. Torgo

ltorgo@dcc.fc.up.pt
Departamento de Ciência de Computadores
Faculdade de Ciências / Universidade do Porto

Sept, 2014

The Problem and its Objectives

- High concentrations of certain harmful algae in rivers is a serious ecological problem with a strong impact not only on river lifeforms, but also on water quality.
- Being able to monitor and perform an early forecast of algae blooms is essential to improve the quality of rivers.
- With this goal several water samples were collected in different European rivers at different times during a period of approximately one year.
- For each water sample, different chemical properties were measured as well as the frequency of occurrence of 7 harmful algae.
- Some other characteristics of the water collection process were also stored such as the season of the year, the river size, and the river speed.

Motivation

- Chemical monitoring is cheap and easily automated, while the biological analysis of the samples to identify the algae is expensive and slow.
- Obtaining models that are able to accurately predict the algae frequencies based on chemical properties would facilitate the creation of cheap and automated systems for monitoring harmful algae blooms.
- Another objective of this study is to provide a better understanding of the factors influencing the algae frequencies.

The Available Data

- There are two main data sets available: one for model development and the other for model testing
- The first contains 200 observations while the second contains 140
- Each observation contains information on 11 descriptive variables: 3 nominal and 8 numeric.
- Each observation is in effect an aggregation of the data on several water samples collected on the same river througout the same season of the year.
- The 3 nominal variables describe the season of the year, the river size, and river speed, for the respective aggregated observation
- The 8 remaining variables describe several aggregated values of chemical parameters measured on the water samples (e.g. maximum pH, minimum value of O_2, etc.)

The Available Data (cont.)

- Associated with these 11 variables there are 7 values of the measured frequency of 7 harmful algae on the respective water samples.
- For the test set (140 observations) no information is given on these 7 variables. Our goal is exactly to forecast these 140×7 values.

The Available Data

- The data sets are available in the DMwR package
- To use the data of the 200 observations it is sufficient to do:

```
library(DMwR)
data(algae)
```

- You may check the first few lines of the data as follows,

```
head(algae)

##   season  size  speed mxPH mnO2    Cl    NO3    NH4   oPO4    PO4 Chla
## 1 winter small medium 8.00  9.8 60.80  6.238 578.00 105.00 170.00 50.0
## 2 spring small medium 8.35  8.0 57.75  1.288 370.00 428.75 558.75  1.3
## 3 autumn small medium 8.10 11.4 40.02  5.330 346.67 125.67 187.06 15.6
## 4 spring small medium 8.07  4.8 77.36  2.302  98.18  61.18 138.70  1.4
## 5 autumn small medium 8.06  9.0 55.35 10.416 233.70  58.22  97.58 10.5
## 6 winter small   high 8.25 13.1 65.75  9.248 430.00  18.25  56.67 28.4
##     a1   a2   a3  a4   a5   a6  a7
## 1  0.0  0.0  0.0 0.0 34.2  8.3 0.0
## 2  1.4  7.6  4.8 1.9  6.7  0.0 2.1
## 3  3.3 53.6  1.9 0.0  0.0  0.0 9.7
## 4  3.1 41.0 18.9 0.0  1.4  0.0 1.4
## 5  9.2  2.9  7.5 0.0  7.5  4.1 1.0
## 6 15.1 14.6  1.4 0.0 22.5 12.6 2.9
```