Knowledge Discovery in Environmental Databases

Data mining and knowledge discovery are concerned with extracting interesting and useful patterns, information, or knowledge from databases. The Department of Intelligent Systems at the Jozef Stefan Institute, headed by Professor Ivan Bratko, has an outstanding expertise in the development and practical applications of data mining methods based on machine learning. These methods include the induction of classification rules, decision trees, and regression trees, inductive logic programming or relational learning, and equation discovery.

The advantages of machine learning methods over classical statistical methods include:

- a richer and more structured form of knowledge discovered, where symbolic and numerical information are combined,
- the ability to take into account existing domain knowledge, and
- the ability to find knowledge that involves more than one relations in a relational database.

Diverse application areas have been addressed within the Department. These include medical diagnosis and prognosis, mechanical engineering, pharmacology, and ecology. Successful applications of data mining and knowledge discovery in environmental databases include:

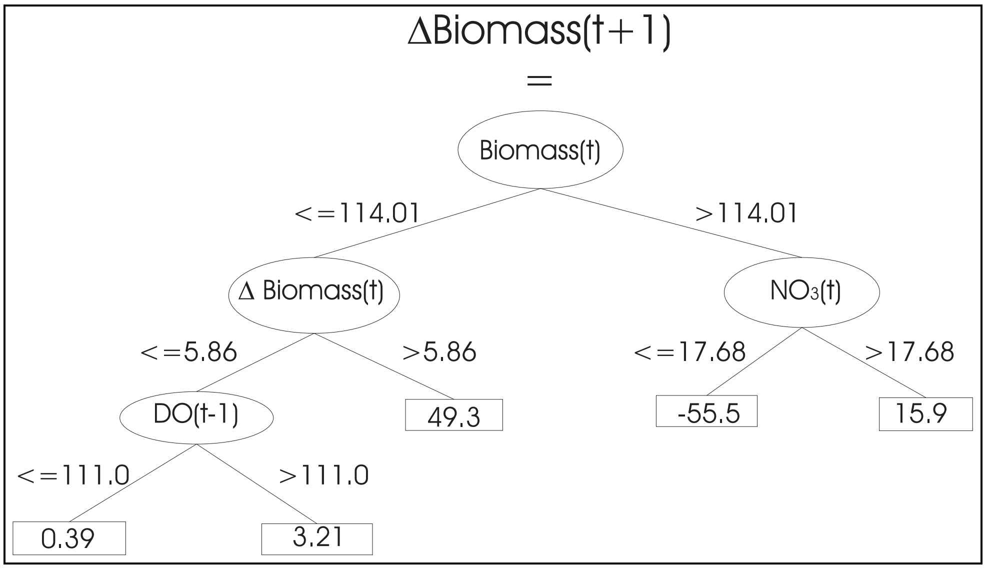
- analysis of biological and chemical data on the water quality of Slovenian rivers,
- biological classification of British rivers,
- modeling algal growth in the Lagoon of Venice,
- modeling algal growth in the Lake of Bled,
- modeling phytoplankton growth in the Danish lake Glumsoe,
- predicting biodegradability from chemical structure, and
- runoff prediction from rainfall and past runoff
  for the River Rott catchment in Bavaria, Germany.

Modeling algal growth in the Lagoon of Venice

The Lagoon of Venice measures 550 km^2, but is very shallow, with an average depth of less than 1m. It is heavily influenced by anthropogenic inflow of nutrients - 7 000 000 kg/year of nitrogen and 1 400 000 kg/year of phosphorus. These loads (mainly nitrogen) are above the Lagoon's admissible trophic limit and generate its dystrophic behavior, which is characterized by excessive growth of algae, mainly *Ulva rigida*.

Samples are collected weekly by the Department of Industrial Chemistry, University of Padova at four different locations in the Lagoon. The sampled quantities are nitrogen in ammonia NH_3, nitrogen in nitrate NO_3, phosphorus in orthophosphate PO_4 (all in μg/l), dissolved oxygen DO (in % of saturation), temperature T (degrees C), and algal biomass B (dry weight in g/m^2).

We applied regression tree induction and equation discovery to model the growth of *Ulva rigida*. An example regression tree is shown in the figure below. It predicts the change of biomass at week $t+1$ ($\Delta Biomass(t+1) = Biomass(t+1) - Biomass(t)$) from the sampled quantities in weeks t and $t-1$ and the change of biomass at week t ($\Delta Biomass(t)$). The tree indicates that the change of biomass at week $t+1$ is most strongly influenced by the biomass at week t, followed by the change of biomass at week t and the nitrogen in nitrate NO_3 at week t, further followed by the DO at week $t-1$. Large decreases in algal biomass can be expected if biomass at week t is higher than 114 g/m^2 dry weight and NO_3 at week t is lower than 17.68 μg/l, while large increases can be expected if biomass at week t is lower than 114 g/m^2 dry weight and the change of biomass at week t is greater than 5.86 g/m^2 dry weight.

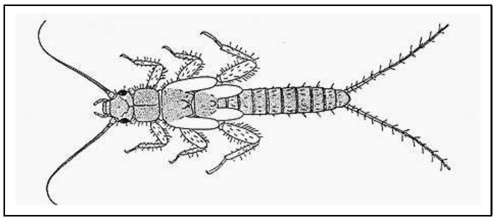


Differential and difference equations that model algal growth were also discovered from the measured data and some basic knowledge on growth and mortality rates. Both the regression tree models and the difference equations perform comparably to expert designed models in terms of prediction. They also provide insight into the relative importance of each of the measured parameters for algal growth.

Analysis of biological and chemical data on the water quality of Slovenian rivers

We have applied machine learning, more specifically techniques for induction of classification rules, to analyze biological and chemical data on the water quality of Slovenian rivers. The data were collected by the Hydrometeorological Institute of Slovenia that regularly performs water quality monitoring for most Slovenian rivers and maintains a database of water quality samples. Both biological and chemical samples were used in the analysis. Three related problems were addressed: analyzing the influence of physical and chemical parameters on selected bioindicator organisms, biological classification of rivers based on physical and chemical parameters, and biological classification of rivers based on bioindicator data. In all three cases, valuable knowledge was extracted from the data acquired through environmental monitoring and expert interpretation of the acquired samples.

In the following, we give some highlights from the analysis of the influence of physical and chemical parameters on selected bioindicator organisms. Using data on the presence/absence of bioindicator taxa in a biological sample and the values of physical and chemical parameters in a corresponding chemical sample, we induced rules that predict the presence or absence of taxa. In this way, we can determine the ecological requirements of organisms that are not sufficiently well understood.



Expert evaluation of the induced rules shows that they capture useful knowledge. In some cases, the rules confirm the expert knowledge of the biology of the bioindicator taxon concerned. For example, the rules that predict the presence/absence of the beetle *Elmis sp.* confirm that it is a bioindicator of clean to mildly polluted waters. Induced rules may also reveal new aspects of the biology of the studied taxon which extend existing knowledge. Consider the two rules below that predict the presence of *Leuctra sp.* (depicted above).

Leuctra sp. (stoneflies) Present
IF Temperature $< 23\ ^\circ$ C
AND 120 % $<$ Oxygen Saturation $<$ 150 %
AND Chemical Oxygen Demand $>$ 10.9 mg O_2/l
AND Biological Oxygen Demand $<$ 3.75 mg O_2/l

Leuctra sp. (stoneflies) Absent
IF Temperature $< 22.25\ ^\circ$ C
AND Total Hardness $<$ 18.55 mg$/l$
AND Biological Oxygen Demand $>$ 6.9 mg O_2/l

The first rule indicates that the taxon is found in clean waters (low COD), which is consistent with existing knowledge about its ecological requirements: *Leuctra sp.* is used as an indicator of clean waters. The second rule, however, indicates that the taxon can be present in quite polluted waters (high COD), provided enough oxygen (high Saturation). This rule supplements existing expert knowledge in an important way as most Slovenian rivers are torrential streams and thus well aerated.

In cooperation with the Center for knowledge transfer in information technologies (Dr. Tanja Urbančič), and the Faculty of Civil Engineering, University of Ljubljana (Dr. Boris Kompare), the Department organizes courses and workshops titled "Analysis of ecological data with machine learning". These courses and workshops are intended for end users and are typically attended by people from industry and governmental institutions that deal with different types of environmental problems, such as water quality monitoring/control and waste management.

The courses give an overview of several machine learning approaches and their applications to mining environmental data. Hands-on exercises are included to provide end users with a perspective of what kind of data mining problems can be addressed using these approaches. The workshops take end users one step further: their databases on environmental problems are mined for knowledge using the machine learning approaches presented at the courses. This enhances the value of the data as the discovered knowledge is a stepping stone for better understanding of environmental problems and processes and improvements in environmental monitoring and management practices.

For more information on knowledge discovery in environmental databases contact

---

Dr. Sašo Džeroski
Department of Intelligent Systems
Jozef Stefan Institute
Jamova 39
SI-1000 Ljubljana
Slovenia

Phone: +386 1 477 3217 or +386 1 477 3900
Fax: +386 1 425 1038 or +386 1 2519 385
Email: Saso.Dzeroski@ijs.si
URL: http://www-ai.ijs.si/SasoDzeroski/SasoDzeroski.html
