Machine Learning of Environmental Spatial Data

Mikhail Kanevski¹, Alexei Pozdnoukhov², Vasily Demyanov³

¹Institute of Geomatics and Analysis of Risk, University of Lausanne
(Mikhail.Kanevski@unil.ch)

²National Centre for Geocomputation, National University of Ireland Maynooth, Ireland

³Institute of Petroleum Engineering, Heriot-Watt University, Edinburgh

WORKING PAPER

There is a growing demand for new adaptive processing tools for different environmental problems: spatio-temporal measurements, satellite images, time series of monitoring, environmental risks and natural hazards assessments, renewable resources estimates, topo-climatic and meteorological measurements, etc. The models and approaches proposed for such problems should be nonlinear, robust and automatic able to work in a changing environment with noisy and variable and several spatio-temporal scales data, capable to integrate science-based models. Very often the input space – the space of independent variables, high dimensional and is composed of geographical coordinates and other relevant features, e.g., generated from digital elevation models. An important task is work and to characterize uncertainties both in original data and in the results.

Such tools can be provided by Machine Learning (ML), which is a general and powerful field for processing and nonlinear universal modelling of complex high dimensional data.

The workshop will present the basic concepts underlying a wide range of conventional ML algorithms and provide the cutting-edge data analysis, modelling and visualisation tools:

- Artificial neural networks: multilayer perceptrons, radial basis function networks, general regression and probabilistic neural networks,
- Self-organizing Kohonen maps,
- Support vector machines and other kernel-based methods.

Real case studies from environmental a variety of problems, like pollution (soil, water systems), climate (temperature and precipitation in a complex regions), natural hazards (landslides, avalanches), renewable resources (wind fields) and other fields of applications will be outlined focusing on the software tools used.

The workshop will be useful both for the beginners and advanced researchers and users.

Some topics of general interest – predictability, complexity, automatic data processing will be presented in detail.

Workshop deliverables: tutorial slides, detailed "how-to-do-it" case studies, software tools, and datasets.
The workshop is based on the following books of the authors:


ACKNOWLEDGMENTS

The authors would like to thank to Dr. V. Timonin, who contributed to the development of machine learning software. Many thanks to colleagues with whom many interesting case studies on environmental data mining were carried out: Dr. D. Tuia, Dr. L. Foresti, G. Matasci, M. Volpi, Ch. Kreis, Ch. Kaiser. Partly the research was carried out within the framework of Swiss National Science Foundation projects “GeoKernels. Phase2” and “KernelCD”.
Lecture 1. Predictive learning from environmental data

MAIN TOPICS:

A. ENVIRONMENTAL DATA:
   a. spatial, temporal, spatio-temporal;
   b. multivariate;
   c. noisy;
   d. extremes and outliers;
   e. multiscale variability;
   f. nonstationarity

Representativity of data? Predictability?

Data = Information(structure, patterns) + noise

B. PREDICTIVE LEARNING:
   a. basic problems and concepts
   b. some theory
   c. model selection and model assessment

C. ILLUSTRATIVE CASE STUDIES
Lecture 2. Spatial prediction of environmental data

2.1 First model: k-Nearest Neighbours = benchmark model

Patterns (right) and corresponding k-NN cross-validation curves

2.2 Multilayer Perceptrons – workhorse of machine learning:

- Data preprocessing
- Construction of MLP model
- Training of MLP model
- Regularization techniques
- Analysis of the results (residuals)
- Simulated case studies
- Real data case studies

Noise injection as a regularization procedure
2.3 General Regression Neural Networks

GRNN estimate at a node $D_i$ from samples $Z_i$:

$$Z_{OUT} = \frac{\sum_{i=1}^{N} Z_i \exp \left(-\frac{D_i^2}{2h^2} \right)}{\sum_{i=1}^{N} \exp \left(-\frac{D_i^2}{2h^2} \right)}$$

GRNN Mapping: detection of patterns, modelling, analysis of the residuals, mapping, assessment of the uncertainty.
Lecture 3. Introduction to Statistical Learning Theory

- Support Vector Machines
- Support Vector Regression

Support Vector Classification/Regression

- Based on Statistical Learning Theory
- Non-linear Robust Classification and Regression
- Kernel Method
- High dimensional, prone to over-fitting
- Allows for training errors
- Unique solution (unlike MLPs)
- Probabilistic interpretation of the classification
- Multiclass classification
- Advanced topics:
  - Active Learning
  - Monitoring networks optimization
  - Multiple kernel learning
Lecture 4. ANNEX topics

4.1 Geostatistics and MLA

4.2 ANNEX Models

4.3 Multitask learning

Manifold learning
4.4 Self-Organizing Maps (Kohonen Maps)

SOM is an algorithm that projects high-dimensional data usually onto a two-dimensional map (rectangular or hexagonal) having \((M_x \times M_y)\) neurons/nodes. Each neuron is associated with a weight vector with the dimension equal to the dimension of the data.

The projection preserves the topology of the data so that similar data will be mapped to nearby locations on the map.

**Self-organizing (Kohonen) map**

**SOM classification of spatio-temporal environmental data**

**Discussions, Conclusions, Future Research**