Input selection and sensitivity analysis of neural networks in hydrology

Jarkko Tikka

tikka@cis.hut.fi

T-61.6060 Special Course in Computer and Information Science VI
Data analysis and environmental informatics
15.2.2006

Contents

Introduction

- Prediction tasks are typical in ecology
- Standard statistical approaches earlier
- Artificial neural networks (ANN) since last decade
- The goal is to estimate function $y = f(x_1, \ldots, x_m)$
- A multilayer perceptron (MLP) network

$$y = \mu + \sum_{j=1}^{p} \beta_j I_j (\sum_{i=1}^{m} w_{ij} x_i + b_j)$$

The main steps in the construction of an ANN model

- data pre-processing
- division of data
- choice of a performance criterion
- selection of model inputs
- determination of network architecture
- training
- testing and sensitivity analysis

- The problem of the MLP networks: black box models
- Unable to clarify actual dependencies
- Especially in ecology, important to evaluate significance of inputs
- Input selection
  - Evaluation of methods for the selection of inputs for an artificial neural network based river model (Bowden, Dandy, Maier)
- Sensitivity analysis
  - Utility of sensitivity analysis by artificial neural network models to study patterns of endemic fish species (Gevrey, Lek, Oberdorff)

Input selection

Advantages:

- increases understanding of the problem
- decreases the number of parameters
- decreases computational complexity
- helps to avoid overfitting

Accomplished in two phases

- unsupervised
- supervised

Unsupervised input selection

A priori knowledge

- to use expert knowledge of the system
- very subjective and case dependent

Self-organizing map (SOM)

- is used to cluster the inputs
- sample one input from each cluster

Principal component analysis (PCA)

- select a few of first principal components
- PCs are independent of each other

Supervised input selection

Genetic algorithm (GA)

- a powerful optimization technique
- Initialization: a population of random solutions is generated
- Evaluation: fitness of each member of the population
- Selection and crossover
- Mutation: small random changes
- Repeat until converges or the maximum number of generations is exceeded

Stepwise procedure

- Add inputs sequentially to the model
- stop when accuracy of the model do not improve

Implementation

- Commercially available software package NeuroGenetic Optimizer
  - the inputs used
  - number of hidden layers
  - the number of neurons in each layer
  - the transfer functions

Case study I

- Objective: to predict (4 weeks in advance) amount of cyanobakteria in the River Murray at Morgan, South Australia
- The most important characteristics: the onset, peak and duration of a bloom
- Dependent variable: concentrations of the cyanobacterium
- Independent variables (10 in total): total phosphorus, soluble phosphorus, total kjedahl nitrogen, silica, turbidity, color, pH, temperature, river levels at Morgan, and weekly flows
- Weekly measurements from 1980/1981 to 1995/1996

Table 1: Results for Case study I.

	a priori knowledge (70)		PCA (48)		SOM (39)	
	GA	Stepwise	GA	Stepwise	GA	Stepwise
inputs	38	20	22	10	25	20
training	256	382	372	439	420	398
validation	505	491	561	578	504	503
test	386	517	436	549	436	602

A priori knowledge + GA gives the minimum test error.
The onset and duration are modeled well and peaks poorly.
Silica and river levels dropped out from the final model.

Sensitivity analysis

- Input selection may not be enough
- The contribution of the inputs to the output is interesting
  - sensitivity analysis
- Many approaches, for instance
  - weights method
  - stepwise method
  - profile method
  - Partial derivatives method

The derivatives profile

$$d_i = \frac{\partial y}{\partial x_i} = S \sum_{j=1}^{p} \beta_j (1 - I_j^2) w_{ij}$$

A graph of the partial derivatives versus each corresponding input variable can be plotted.

The relative contributions (Sensitivity for a set of data (SSD))

$$SSD_i = \sum_{j=1}^{N} d_{i,j}^2, \qquad SSD_i \leftarrow \frac{SSD_i}{\sum_{i=1}^{m} SSD_i}$$

The sensitivity of the output y for the data set with respect to an input x_i

Case study II

- Data were collected from 136 rivers of the Northern Hemisphere
- Objective: to predict endemic species richness (ESR)
- Independent variables (3 in total): total species richness (TSR), the total surface area of the drainage basin (SAD), and net primary productivity (NPP)
- Average values of NPP were calculated using mean annual air temperature and the mean annual rainfall.
- MLP network, architecture 3-5-1

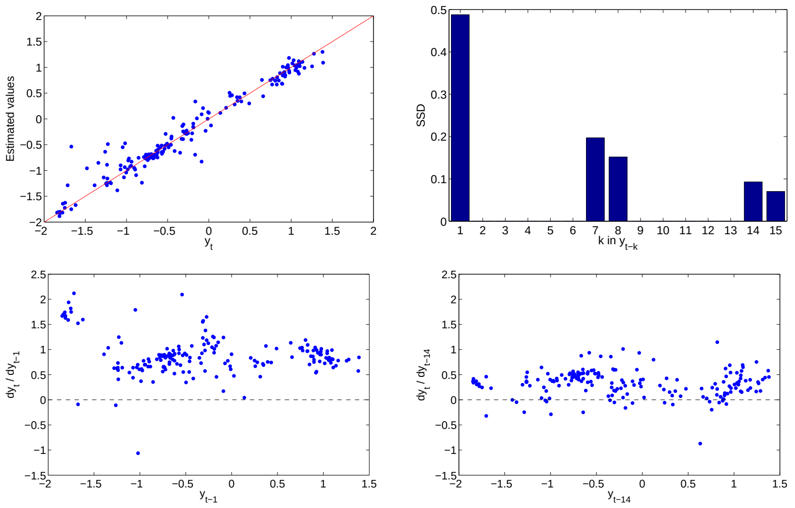
- coefficient of determination is 0.92.
- Partial derivatives:

$\partial ESR/\partial TSR$: with small values of TSR nearly zero and with larger values of TSR positive
$\partial ESR/\partial SAD$ and $\partial ESR/\partial NPP$: nearly zero with all the values of SAD and ESR, respectively

- Relative contributions: TSR 77%, SAD 20%, and NPP 3%
- Without SAD and NPP the coefficient of determination is 0.83

An additional example

- Data: daily measurements from the electricity consumption
- Objective: to predict electricity consumption
- One-step-ahead prediction: $y_t = f(y_{t-1}, \dots, y_{t-15})$
- Inputs are selected based on linear model
- Final inputs: $y_{t-1}, y_{t-7}, y_{t-8}, y_{t-14}$, and y_{t-15}
- 5 times 10-fold cross-validation
- MLP architecture: 5-7-1
- MSE for the test set: 0.038



Estimated values
y_t
SSD
k in y_{t-k}
dy_t / dy_{t-1}
y_{t-1}
dy_t / dy_{t-14}
y_{t-14}