Visualisation of variable importance in neural network regression and classification

Kucheryavskiy, Sergey

Published in:
XIII International Conference on Chemometrics in Analytical Chemistry

Publication date:
2012

Document Version
Early version, also known as pre-print

VISUALISATION OF VARIABLE IMPORTANCE IN NEURAL NETWORK REGRESSION AND CLASSIFICATION

Sergey Kucheryavskiy

Department of Biotechnology, Chemistry and Environmental Engineering,
Aalborg University, Niels Bohrs vej 8, 6700, Esbjerg, Denmark
svk@bio.aau.dk

Artificial Neural Networks (ANN) is an efficient tool for solving many machine learning problems, from exploratory analysis to classification and regression [1]. Being quite popular in late 90s, now it is not very widespread in chemometric applications. One of the reasons, from our point of view, is that ANN does not provide enough visual information about model itself as well as about influence of particular variables and samples on the model. Thus, for example, in PLS there are several plots giving an information about importance of variables: weights, regression coefficients, VIP scores and selectivity ratio [2], while ANN tools do not have such option. In this work an attempt to solve this problem was made.

Any neural network can be represented as set of interconnected nodes (neurons). Every node has several inputs $\{x_1,\ldots,x_n\}$ and one output, y. The output value depends on a weighted sum of input values and an activation function F. If we connect several nodes to a simple network, the output of the whole net will depend on a composition of all nodes weights and functions.

Training NN model means finding optimal values for the weights that minimize error of prediction of the output value. One of the most popular training algorithm — backpropagation — implies that we find an overall cost function for the whole network first and then minimize it iteratively. To calculate the cost function the prediction error is propagated from the output back to the inputs.

The general idea going to be presented in this work is to use similar way to propagate weights of all nodes inputs for calculating an importance of the input variables x. The importance of a particular variable is computed as a combination of weights of all nodes, linked with the variable. The method has been applied for analysis of spectral data using autoassociative network for exploratory analysis (analogue of PCA) and multilayer perceptron for regression and classification. Different schemes of weight combinations were tried to find the optimal one. The obtained results show that the proposed method gives a decent illustration of the variable importance and can be also used for variable selection purposes.

References:

[1] F. Marini, R. Bucci, A.L. Magri, A.D. Magri, Microchemical Journal, vol. 88 (2), 178 (2008).
[2] C.M. Andersen, R. Bro, J. of Chemometrics, vol. 24, 728 (2010).