FOSS

A White Paper from FOSS

Artificial Neural Networks and Near Infrared Spectroscopy - A case study on protein content in whole wheat grain

By Lars Nørgaard*, PhD, Senior Manager Team Chemometric Development, FOSS
& Affiliated Professor, University of Copenhagen
Martin Lagerholm, Chemometrician, FOSS
Mark Westerhaus, Research Scientist, FOSS

*corresponding author lno@foss.dk

Issue 1, April 2013

Dedicated Analytical Solutions

Artificial Neural Networks and Near Infrared Spectroscopy - A case study on protein content in whole wheat grain

Artificial Neural Networks are well-established calibration methods with explicit advantages when modelling large and complex databases. The basic principle of ANN will be presented and the power of ANN exemplified by a case study on prediction of protein content in whole wheat grain by NIR.

By Lars Nørgaard*, Martin Lagerholm and Mark Westerhaus, FOSS
*corresponding author lno@foss.dk

Introduction
The concept of Artificial Neural Networks (ANN) has provided solutions and insight into a multitude of complex problems since its introduction as a powerful data analytical tool in the 1980-1990'ies [1]. The efficiency of ANN methods is undisputed but the methods are relatively complex when it comes to implementation, method setup, training and estimation of parameters compared to e.g. linear regression methods as Principal Component Regression or Partial Least Squares Regression. The last-mentioned methods are widely used for application development relating spectroscopic data to relevant reference analyses while the use of ANN methods has been more restricted. Here we will describe the principle of ANN and apply the method for the development of global calibration models on large databases - the case studied is the global NIR Infratec prediction of the concentration of protein in whole wheat grain [2].

From the brain to the computer
The original inspiration for the development of artificial neural networks came from neuroscience. The neurons in the brain are connected through complex networks and this concept was the inspiration for the Artificial Neural Network as an analogy to the human biological neural network. The human brain possesses extraordinary pattern recognition abilities with regards to the world around us; based on sensory input transmitted to the brain, decisions are taken and patterns are discerned and recognized with a high level of confidence. On the other hand it is very difficult for the human brain to extract qualitative and especially quantitative information regarding protein content in whole wheat grain from NIR spectra! In Figure 1 NIR spectra of 100 randomly selected whole wheat grain samples are shown. The spectra are colored according to protein content and it is obvious that it is difficult for most human brains to extract quantitative and accurate information - in such cases the Artificial Neural Networks will do an excellent job.

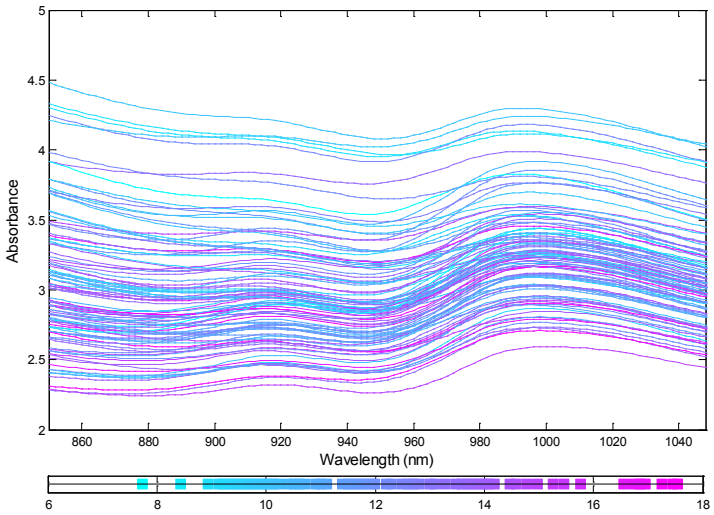

Figure 1. One hundred randomly selected NIR spectra of whole wheat grain samples colored according to protein content from 7.74 % to 17.54 %.

Network design

A neural network can have a multitude of designs; as an example we will focus on the so-called feed forward two layer network that includes a) inputs, b) a hidden layer and c) an output layer - see Figure 2 (inputs do not count as layers in ANN terminology). The input neurons - first layer of neurons or units - simply represent the recorded spectral values; in the case of NIR transmission spectra of whole wheat grains we have recordings for every 2nd nm from 850 nm to 1048 nm corresponding to 100 input data. Different pre-processing methods can be applied prior to feeding the spectra into the network; we will not describe these here but stress that intelligent pre-processing is essential to obtain robust and well-performing models; at FOSS we use a proprietary mathematical preprocessing stage before artificial neural network calibrations are developed.

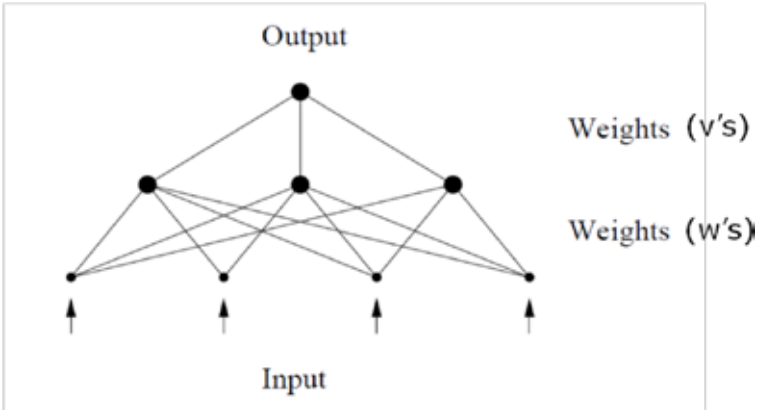

Figure 2. Simplistic feed forward ANN where each circle represents a neuron; in this example three hidden neurons are used and only four inputs (representing spectral variables) with one output (protein). All connections are directed from the input towards the output [3].

The input data points - e.g. spectral or pre-processed measurements - are combined by weights exactly as in a linear regression model:

x1*w11 + x1*w12 + … + x100*w1100 = ihn1
x1*w21 + x1*w22 + … + x100*w2100 = ihn2
x1*w31 + x1*w32 + … + x100*w3100 = ihn3

where ihn1 means input to hidden neuron number 1 and w2100 superscript means 2nd hidden neuron while 100 as the subscript means weight number 100 . The non-linearity is normally dealt with in the hidden layer between the input data (spectra) and the output layer (protein). Suppose we have three hidden neurons in our setup; then we will calculate ihn1, ihn2 and ihn3. Each of these will be transformed through a non-linear function which is the key to the non-linear modeling of an ANN model; by selection of a proper non-linear function the degree of non-linearity needed will be regulated by this transfer function. In Figure 3 an example of a sigmoid non-linear transfer function is shown; a multitude of non-linear transfer functions can be used with different purposes to provide optimal solutions on specific applications.

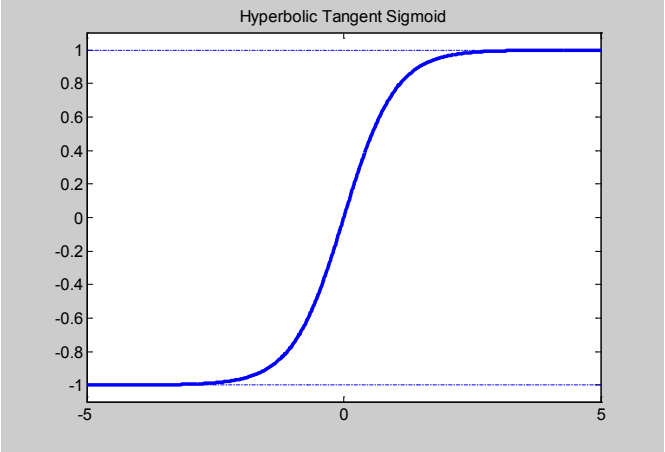

Figure 3. A non-linear transfer function - the hyperbolic tangent sigmoid function.

The output from the hidden neurons – ohn1 to ohn3 – is then linearly weighted to provide a joint "input to the output layer"
ohn1*v1 + ohn2*v2 + ohn3*v3 = "input to output layer"
This weighted sum can go through a linear function or a non-linear function to increase further the flexibility in the non-linear modeling.
Normally, biases are also included in the above equations but for simplicity we have left out these in the description.

Whole wheat grain protein quantification with ANN
Let's now develop an ANN prediction model for protein in whole wheat grains. The data are from the FOSS wheat and barley database.

Database & data sets
It is important to have a wide range in protein and moisture values and to obtain that, harvest data has been collected for 25 years straight and today the FOSS database includes more than 50.000 samples. In addition 25 years of broad seasonal, geographical, grain type, instrument, spectroscopic scatter and temperature variations are included. It is extremely important to have as much relevant variation represented in your cali-

bration database in order to obtain robust and accurate predictions as well to secure stability over time.
In the present case we will focus on protein and the range covered is from 7-24% protein. The calibration data set has the following properties [4]

- More than 40 000 wheat samples
- Samples have been collected globally over 25 years
- Temperature variation of 0-40 °C has been added in a controlled way
- Many different varieties are included
- Instrument variations from Infratecs of all different generations

In order to estimate the performance of the global ANN calibration model, predictions of an independent global validation set are evaluated. The validation set consists of 11908 samples.

Model development & prediction
An ANN is trained on more than 40.000 samples from the database to obtain the optimal setting of all the weights in the network - as described this is a complex task but it is easily taken care of by the computer. When the ANN is trained the weights are fixed and we have an ANN model where we know all the settings and it is now easy to perform the protein prediction for a new sample. The spectrum is recorded (50 seconds) and input to the ANN architecture with the trained weights. By pre-processing, multiplication, addition and non-linear function transfers of these input data, we obtain an estimate of the protein content of the sample analysed. In Figure 4 the model is evaluated on the independent test set covering all possible variations. The RMSEP is 0.27 % and the bias is 0.011 % and the model is very robust with respect to seasons, regions, temperatures, instruments and grain types. Linear models are not able to deal with these non-linear data and will provide a much poorer performance [2].

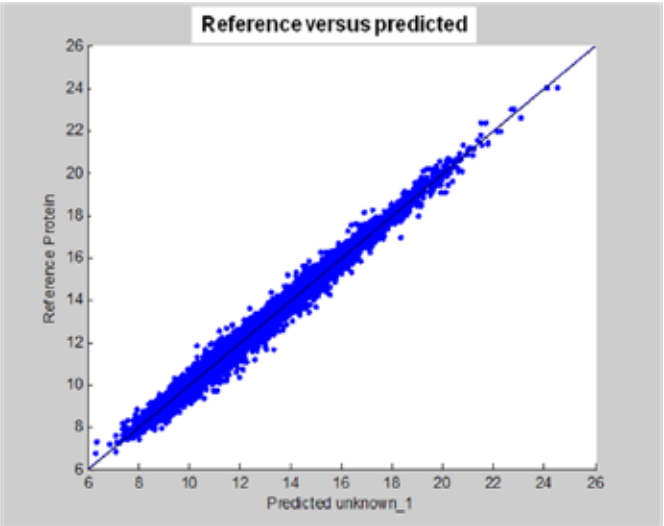

Figure 4. The independent test set prediction compared to the reference values as obtained by wet chemistry.

As has been shown in a previous Chemometric Corner, the combination of an Infratec, an ANN calibration and a comprehensive database outperforms the accuracy of the corresponding laboratory analysis method [5].

Training of ANN models

For linear models it is a quite straightforward task to estimate the regression coefficients e.g. through a least squares estimation, minimizing the residuals between the laboratory reference values and the predicted values. In a linear model we estimate 100 regression coefficients plus an offset - in total 101 coefficients. In ANN modeling the estimation part is more complex; in the case presented we need 100 x 3 neurons (w's) + 3 x1 (v's) (excluding biases) besides the complexity in estimating the weights over several layers and non-linear functions. This is the price to be paid for having a model capable of handling non-linearity, and the estimation of the coefficients can be performed by optimization methods from numerical analysis; e.g. by gradient descent or more advanced modern and tailored methods.

When training a network the goal is to get the network to respond to a given input spectrum in a clever way. There are two states: a) a training or learning state where the ANN learns how to respond and b) a prediction state where the network is applied to a spectrum it has never seen before. In this prediction state no corrections to the weights are made. The training and learning takes place by comparing the protein output from the network with the reference protein value and then changing the weights to get closer to the reference value. So the network learns from example just as the brain could do. The learning occurs simply by changing the weights in a systematic way.

Outro

Artificial neural networks are very efficient for the extraction of quantitative information from large spectroscopic databases where non-linearity is inherent due to complex biological, environmental and instrumental variations. The basis for a well-performing ANN model is a comprehensive database spanning all relevant variation and efficient and optimized ANN models and modeling tools. The combination of ANN and a well-equipped database offers improved robustness, stability and last, but not least, accuracy.

References

[1] CM Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995

[2] NB Büchmann, H Josefsson, IA Cowe, Performance of European Artificial Neural Network (ANN) Calibrations for Moisture and Protein in Cereals Using the Danish Near-Infrared Transmission (NIT) Network, Cereal Chemistry, 78(5), 572–577, 2001.

[3] M Lagerholm, Resource Allocation with Potts Mean Field Neural Network Techniques, PhD thesis, Lund University, 1998.

[4] T Nilsson, Comparison of the FOSS NIR global ANN calibration against reference methods: A five year pan-European study, FOSS White Paper, http://www.foss.dk/-/media/Files/Documents/IndustrySolution/Papers/Grain/PanEuroStudy.pdf.ashx, October 2011.

[5] R Malm, M Westerhaus, L Nørgaard, Can NIR & Chemometrics be more accurate than the reference method? A review of global ANN calibrations for whole grain analysis, InFocus 2012, Issue 2.

FOSS
Foss Allé 1
DK-3400 Hilleroed
Denmark

Tel.: +45 7010 3370
Fax: +45 7010 3371

info@foss.dk
www.foss.dk