NEURO-FUZZY MODELING FOR CROP YIELD PREDICTION

D. Stathakis[a*], I. Savin[a] and T. Nègre[a]

[a]EC-JRC/MARS-FOOD, TP266, 21020 (VA), Italy, demetris.stathakis@jrc.it, http://agrifish.jrc.it/marsfood/

ABSTRACT

The purpose of this paper is to explore the dynamics of neural networks in forecasting crop (wheat) yield using remote sensing and other data. We use the Adaptive Neuro-Fuzzy Inference System (ANFIS). The input to ANFIS are several parameters derived from the crop growth simulation model (CGMS) including soil moisture content, above ground biomass, and storage organs biomass. In addition we use remote sensing information in the form of the Normalized Difference Vegetation Index (NDVI). ANFIS has only one output node, the yield. In other words a single number is sought. An additional difficulty in predicting yield is that remote sensing data do not go long back in time. Hence any predicting effort is forced to use a very limited number of past years in order to construct a model to forecast future values. The system is trained by leaving one year out and using all the other data. We then evaluate the deviation of our estimate compared to the yield of the year that is left out. The procedure is applied to all the years and the average forecasting accuracy is given.

1. INTRODUCTION

The synergy of neural network, fuzzy sets and genetic algorithms is collectively known as computational intelligence (Duch et al., 2004). Several different combinations exist. There exist the optimisation of neural network weights via a genetic algorithm (Liu et al., 2002; Liu et al., 2004) and the use of a genetic algorithm to tune a fuzzy inference system (Russo 2000). In this paper we focus on the merge of neural networks and fuzzy set theory which result to fuzzy neural networks (FNN) also known as neuro-fuzzy systems or granular neural networks (GNN) (Pedrydcz et al., 2001).

Several flavours of neuro-fuzzy systems have been applied to the classification of remotely sensed imagery. A comparison of the most commonly used ones can be found in Stathakis et. al. (in press a). We briefly mention here that NEFCLASS (Nauck, 2003) has been used in the classification of remotely sensed data to obtain land use and land cover classes (Stathakis et al., in press b). ANFIS (Jang, 1993) has been used in the same context (Benediktsson et al., 1999). Fuzzy ARMAP (Carpenter et al, 1992) has been applied into global land cover mapping (Gopal et al., 1999).

The use of a neuro-fuzzy system for crop yield estimate has several appealing characteristics. All the variables that are input into the system are associated with varying degrees of accuracy. The ambiguity steams from measurement error and generalisation. Using fuzzy sets instead of the actual values as inputs, we aim at shifting to the semantics of the data rather than its measure (Zadeh, 1999).

The purpose of this paper is to apply ANFIS neuro-fuzzy architecture for crop yield estimation. We proceed by briefly presenting the architecture of this neuro-fuzzy system. We then describe the data and parameters used as well as the exact methodology. Finally conclusions and guidelines for future work are given.

2. METHOD

2.1 ANFIS

We considered several architectures to use in crop yield prediction. Perhaps the most important parameter for selecting the most suitable one is that we wanted that the estimate would be a number. It is well know that with neuro – fuzzy modelling there is the alternative to use a fuzzy set as the output. In this case, yield would be expressed for example as low, normal of high with each of those three hedges corresponding to a fuzzy set. We do not imply however that seeking a crisp (non-fuzzy) value is a more exact approach than seeking a trend expressed in fuzzy sets (low, medium, etc). But although the accuracy of prediction is probably the same in both expressions of desired output, people are more used to and feel more confident in looking at number rather than a membership function.

ANFIS is a system that accepts numerical inputs and produces a single output value. While this is clearly a limitation when using this system e.g. in land use classification, because only one output class is permitable, it is proved here to be an advantage. The multiple output counterpart of ANFIS called CANFIS has also been suggested (Mizutani 1997).

ANFIS is susceptible to the "curse of dimensionality". The training time increases exponentially with respect to the number of fuzzy sets per input variable used. To illustrate this let us consider a system with 8 input variables that are coded into two fuzzy sets (e.g. low, high) and has 256 rules, as shown on figure 1. If we chose now to use three linguistic variables instead of two (say low, medium, high) the number of fuzzy rules becomes 6561. This phenomenon limits in practice the choice of input variables as well as the expression of those variables into meaningful fuzzy sets.
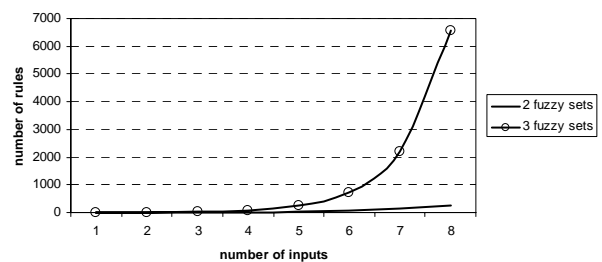

FIG 1. The curse of dimensionality in ANFIS

2.2 Data

Three adjacent regions in Russia (Rostov, Krasnodar, and Stavropol) with similar agro-ecological and agronomical conditions are used as pilot areas for the study. These regions make the biggest contribution in wheat grain production of Russia. Statistical winter wheat data for 1999-2004 were used for the network training. The study area is shown in Fig 2.

Fig. 2. Pilot area location (hatched regions)

2.3 Parameters used

A model for predicting winter wheat yield at regional level is constructed. The input to the network are several parameters derived from the crop growth simulation model (CGMS) (Supit et al., 1994) including soil moisture content, above ground biomass, and storage organs biomass. In addition, we use remote sensing information in the form of the Normalized Difference Vegetation Index (NDVI).

The CGMS outputs and NDVI values originally were received on 10-daily basis during each vegetative season. However, all input parameters were aggregated for two periods to construct the time profiles. The duration of the first period is from crop sowing to maturing (up to the 1st of August). The second period corresponds to the crop development from sowing to flowering (up to the 1st of June). The main reason of such an aggregation is to investigate a difference between "early" (during crop flowering) and "late" (during crop maturing) crop yield prediction.

For each parameter we wither use the maximum value or the first two coefficients of the 5-order polynomials. Those polynomials are used to approximate each parameter. As a result, the following set of input parameters was created:

1. maximum NDVI value for the 2nd period;
2. first and second coefficients of the NDVI 5-order polynomial for the 1st period;
3. first and second coefficients of the NDVI 5-order polynomial for the 2nd period (fig.3);
4. maximum storage organs' biomass value for the 1st period;
5. maximum storage organs' biomass value for the 2nd period;
6. first and second coefficients of the storage organs biomass 5-order polynomial for the 1st period;
7. first and second coefficients of the storage organs biomass 5-order polynomial for the 2nd period;
8. maximum above ground biomass value for the 1st period;
9. maximum above ground biomass value for the 2nd period;
10. first and second coefficients of the storage organs biomass 5-order polynomial for the 1st period;
11. first and second coefficients of the storage organs biomass 5-order polynomial for the 2nd period;
12. first and second coefficients of the soil moisture 5-order polynomial for the 1st period;
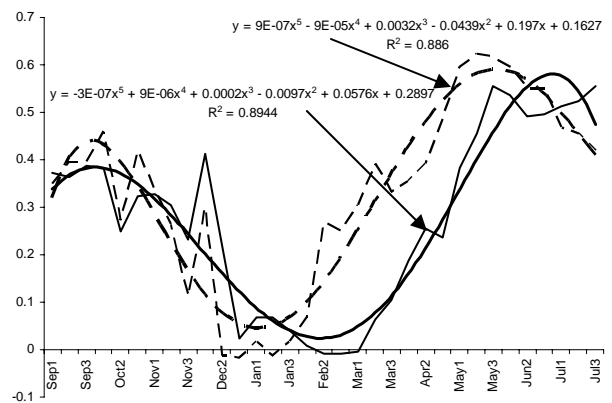13. first and second coefficients of the soil moisture 5-order polynomial for the 2nd period.

Fig 3. Example of NDVI time profiles for two years for Krasnodar region and corresponding 5-order polynomials for the first period of analysis (late period).

2.4 Methodology

The total number of samples is 18. This number is very low and it shows one of the major difficulties in our test. One of the main questions that need to be addressed in this study is whether it is possible which such a low number of training samples to predict future values. The positive aspect compared to statistical methods is that, the neuro-fuzzy method has no conceptual limitation because it is purely data driven. Although not recommended in general, no assumptions are violated by having such a low population number.

ANFIS architecture is determined by preliminary testing. Finally the following parameters were selected for the first neuro-fuzzy configuration: 1, 3, 5, 9, 11 (see above). Parameters 2, 4, 8, 10, 12 were selected for the second one. Triangular membership function is deployed. The stopping criterion is either completion of ten training epochs or reaching 0.01 mean classification error for the training data set. The input parameters are normalized to the [-1, 1] range according to the formula:

$$output = \frac{input - \dfrac{min + max}{2}}{\dfrac{max - min}{2}}$$

ANFIS is very stable. The results remain the same in subsequent runs with the same input. On the contrary we tested standard neural networks that proved to be quite unstable. We tested a feed-forward multi layer perceptron trained by the

backpropagation algorithm (Rumelhart et al., 1986) and we had to average the results over many runs (10-100 times) in order to obtain a stable value.

3. RESULTS

samples	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	overall
period 1	37.5	37.6	31.5	16.7	23.8	36.0	38.2	37.3	56.8	50.8	11.3	34.8	4.0	31.2	31.9	24.1	22.7	23.5	
period 2	24.4	27.0	39.5	25.5	17.5	30.7	32.4	33.8	33.4	30.4	13.7	35.5	43.1	23.6	35.0	25.6	18.4	23.5	
actual	24.4	32.0	28.6	33.1	24.0	34.9	38.2	40.4	44.4	47.5	33.8	43.4	22.3	21.1	30.4	30.7	19.6	35.9	
period 1 correct	46%	82%	90%	50%	99%	97%	100%	92%	72%	93%	33%	80%	18%	52%	95%	78%	84%	66%	74%
period 2 correct	100%	84%	62%	77%	73%	88%	85%	84%	75%	64%	41%	82%	7%	88%	85%	83%	94%	65%	74%

Table 1. Results. The first row gives the ID of the sample. "Period 1" and "Period 2" rows shows the results for late and early periods accordingly. "Actual" row is the actual yield realised at each sample. "Period 1 correct" and "Period 2 correct" is the ratio of estimate divided by the actual yield per sample for the two periods. The last column shows overall percentage of yield estimate.

The results are shown on table 1. We can see that on average the prediction gives a 74% level of accuracy which we consider satisfactory as a first approach, given the very limited number of training samples. In addition, the overall percentages of estimation accuracy for the two periods used are equal (by chance). However, accuracy values for the early period are more stable than for the late period. This can be explained by the fact that the crop yield in the region depends more on the crop growth conditions before flowering than on conditions after flowering. This means that a quite accurate estimate can be made relatively early in time (about two months before harvesting). The internal distribution of error per period exhibits very heterogeneous behaviour. Some samples are estimated very accurately whereas in certain cases the forecasting fails. Different samples are difficult to forecast between the two periods. The graphs of prediction accuracy for periods 1 and 2 are given in figures 4 and 5 accordingly. The structure of ANFIS that we used is shown in fig. 6.
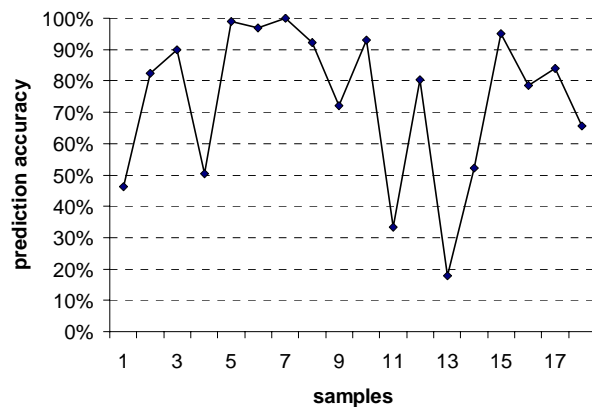

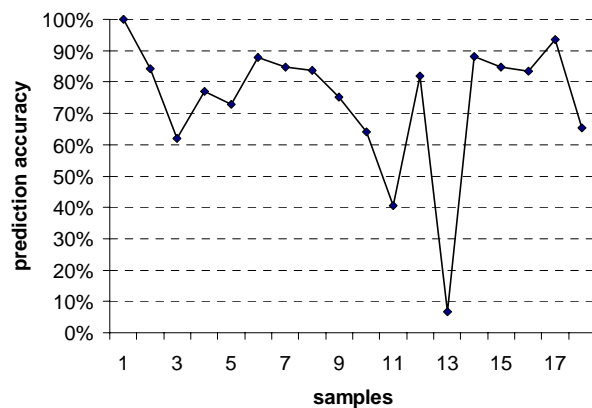Fig 4. Prediction accuracy for the late period (1).

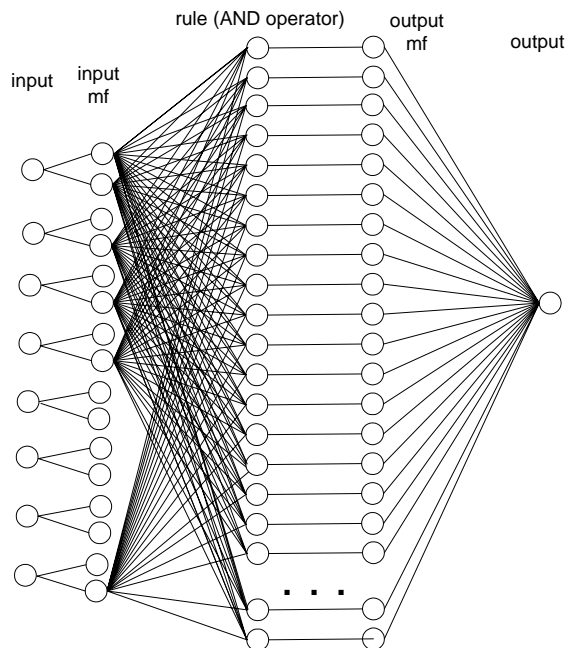Fig 5. Prediction accuracy for the early period (2).

Fig 6. ANFIS architecture for period 2 (early). Not all of the connections and nodes are shown.

4. CONCLUSIONS

This first effort showed that a neuro-fuzzy configuration can be used for wheat yield prediction for the region with acceptable results. In the future we intent to make two additional experiments with the same basic configuration. First use different parameters as inputs and second test more rigorously different ANFIS configurations.

The first group of tests is to find the best input dimensions for this particular task. There is a range of possible parameters other than those used here that we could utilise:

- results of dry matter production modelling based on Monteith approach (Monteith, 1972);
- meteorological parameters (temperature, amount of precipitation, snow depth and so on).

The second group of tests refers to ANFIS parameters. Although we perform a basic sensitivity analysis on the basic parameters used we would be more confident by a more systematic approach. We believe that parameters such as the number of fuzzy sets, the type of membership functions as well as the option to have different parameters per input, should receive more experimental effort. A possible way forward is to use a genetic algorithm to select the optimal values.

Finally, it would be very interesting to explore the knowledge learned by the fuzzy system by looking thoroughly into the rules created in the rule base. Perhaps this would provide an insight with respect to which are the most important factors in crop yield prediction using a data driven approach.

5. REFERENCES

Benediktsson J. A. and H. Benediktsson, Neuro-fuzzy and soft computing in classification of remote sensing data, 1999, SPIE, vol. 3871, *EUROPTO Conference on image and Signal processing for remote sensing V*, Florence, Italy, pp. 176-187.

Carpenter G., Grossberg S., Markuzon N., Reynolds J. and D. Rosen, 1992, Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analogue multidimentional maps, IEEE Trans. Neural Networks, 3(5), pp. 698-713.

Duch, W. and Mandziuk, J., 2004, Quo Vadis Computational Intelligence?, Advances in Fuzzy Systems - Applications and Theory, World Scientific, pp. 3-28

Gopal S., Woodcock C. and A. Stahler, 1999, Fuzzy neural network classification of global land cover from 1o AVHRR data set, *Remote Sensing of Environment*, 67, pp. 230-243.

Jang J.-S. R., 1993, ANFIS: Adaptive-Network-based Fuzzy Inference Systems, *IEEE Trans. Sys, Man, and Cybern.*, 23(3), pp. 665-685.

Liu Z. J., Wang C. Y., Niu Z. and A. X. Liu, 2002, Evolving multi-spectral neural network classifier using a genetic algorithm, IGARSS.

Liu Z., Liu A., Wang C. and Z. Niu, 2004, Evolving neural network using a real coded genetic algorithm (GA) for multispectral image classification, *Future Generation Computer Systems*, 20, p. 1119-1129.

Mizutani E., 1997, Coactive neuro-fuzzy modeling: Toward generalized ANFIS, in Neurofuzzy and soft computing, J-R Jang, Prentice Hall, NJ, chap. 13.

Monteith, J.L., 1972. *Solar radiation and productivity in tropical ecosystems.* J. Applied Ecology, 19:747-766.

Nauck D., 2003, Fuzzy data analysis with NEFCLASS, *International Journal of Approximate Reasoning*, 32, pp. 103–130.

Pedrycz W. and G. Vulkovich, 2001, Granular neural networks, *Neurocomputing*, vol. 36, pp. 205-224.

Rumelhart D.E., J. McClelland and the PDP research group, 1986, Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1- Foundations, MIT, pp. 547.

Russo M., 2000, FuGeNeSys a fuzzy genetic neural system for fuzzy modelling, IEEE Transaction on Fuzzy Systems, 6, pp. 373 - 388.

Stathakis D. and A. Vasilakos, in press (a), Satellite image classification using granular neural networks, *International Journal of Remote Sensing*.

Stathakis D. and A. Vasilakos, in press (b), A comparison of several computational intelligence based classification techniques for remotely sensed optical image classification, *IEEE Transactions in Geoscience and Remote Sensing*.

Supit I., Hooijer A., van Diepen C. 1994. *System description of the WOFOST 6.0 crop growth simulation model.* - JRC, Brussel, Luxembourg, 1994. - 125p.

Zadeh L.A. 1999, From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions, IEEE Transactions on circuits and systems – I: Fundamental theory and application, vol. 45, no. 1, pp. 105-119.