Hydrol. Earth Syst. Sci. Discuss., 3, S186–S189, 2006
www.copernicus.org/EGU/hess/hessd/3/S186/
European Geosciences Union

Hydrology and Earth System Sciences Discussions

Interactive comment on "Optimising training data for ANNs with Genetic Algorithms" *by* R. G. Kamp and H. H. G. Savenije

R. G. Kamp and H. H. G. Savenije

Received and published: 28 April 2006

First of all we want to thank the referee for his valuable comments and the attention given to this paper.

2-Nowadays a lot of hydraulic data is available, however when it comes to a specific project the data seldom fits the project goal. The time period of the available data for example does not fit with the simulation period.

3-I understand that the selection and definition of the subsets are not clear. The selection criteria and the role of the genetic algorithm require more explanation.

The genetic algorithm (GA) selects five random time periods from the original training data and puts the corresponding input and output values in separate subsets. The optimised training set consists of simulation data from a computer model based on a real water system. Both the start and end date of each subset is randomly chosen, which results in a varying length of the subsets. In this paper the GA optimises the

subsets by changing these start and end dates. These five new subsets together form a new training data set resulting in a better-trained ANN. Only small artificial effects occur at the boundaries of these subsets resulting in sharp edges.

It is possible that a given observation is used more than once in the training set. The GA is free to choose the same begin and end date for two different subsets. The number of subsets is arbitrary.

In the final paper we shall make this clear and pay more attention to the role of the GA in the selection of the subsets.

4-For this specific flow model we used two previous time steps in the stepped delay line. The 2nd and the 7th time step were used and were estimated by the cross correlation of the input and output data. The selection of the delay line corresponds to the characteristics of the flow model. Important parameters are for example the length of the flow channel and the shape of the cross sectional flow profile.

5-In a traditional approach, one data set is available for both training and testing. An additional cross validation data set can be used to prevent over-fitting. The problem is that the test set can for example be too small or badly selected and does not show the ANN's capability to generalise new data. For the particular case it is possible to generate simulation data with the flow model. In this paper not one but three test sets were used to test the trained ANN on it's generalisation capacity. The test sets were constructed from model simulation data and had different characteristics. With "differently shaped" on page 289 (section 4) it was stressed that the test sets did not have the same data source. In the final paper we will pay more attention to this testing procedure.

6-I realise we did not fully explain the method we used in this paper. The introduction says there are many reasons for the lack of data. The introduction also shows that ANNs can simulate or mimic hydrological and hydraulic flow models. ANNs need enough training data to perform an appropriate training set to use for example the

HESSD

3, S186–S189, 2006

Levenberg-Marquardt backpropagation training algorithm. In this paper it should be stressed that because of the lack of data a totally new data set was constructed within the limits of the model's hydraulic constrictions. The construction of this new data set needs more explanation.

This new data set is an artificial data set constructed with the existing hydrological flow model. The advantage is that the flow model can easily make new input-output data sets. The disadvantage of this approach is that one does not know how long and what shape the data set must have to become a good training data set. The assumption made was that the discharge must lie between the discharge boundaries and that the data set must have enough variations, increasing from low frequencies (period of days) to high frequencies (period of hours). This however resulted in very long data sets and poor training results. The hypothesis was that a) all available model information lies in this data set and b) the GA is able to select time frames in this particular subset data set which are more useful and effective in the training phase than others.

Summarising, in most cases sufficient data is not available especially if one wants to simulate or mimic a flow model with an ANN. It is possible to construct artificial data sets that have all necessary model information. However, these data sets are very long and produces poor results when trained with an ANN. The GA is used to decrease the size of the data set and at the same time to improve the ANN performance.

7-It is possible to construct graphs of the original and the optimised data set (see attachment). The initial data set puts emphasis on waves of half a day. The optimised data set has corrected this and put more emphasis on a broader spectrum with wave periods from 2.7 days to 14 hours. Wave periods longer than 2.7 days (left part of graph) were excluded in the optimised data set. The GA created a more balanced training set. We shall include this graph in the final paper.

Fig.4 in the original paper shows an example of an optimised data set.

Fig.5 in the original should have expressed that the five subsets are equally distributed

HESSD

3, S186–S189, 2006

after optimisation by the GA. With this graph one can check if the GA actual had worked. If the GA had not optimised it could have resulted in an irregular distribution graph. This graph can be replaced.

Interactive comment on Hydrology and Earth System Sciences Discussions, 3, 285, 2006.