Hydrol. Earth Syst. Sci. Discuss., 3, 285–297, 2006
www.hydrol-earth-syst-sci-discuss.net/3/285/2006/

Hydrology and
Earth System
Sciences
Discussions

Optimising training data for ANNs with Genetic Algorithms

R. G. Kamp[1,2] and H. H. G. Savenije[1]

[1]Section of Water Resources, Delft University of Technology, Delft, The Netherlands
[2]MX.*Systems* B.V., Rijswijk, The Netherlands

Received: 11 January 2006 – Accepted: 18 January 2006 – Published: 9 March 2006

Correspondence to: R. G. Kamp (robert.kamp@mx-groep.nl)

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◂◂ | ▸▸

◂ | ▸

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Abstract

Artificial Neural Networks have proven to be good modelling tools in hydrology for rainfall-runoff modelling and hydraulic flow modelling. Representative data sets are necessary for the training phase in which the ANN learns the model's input-output relations. Good and representative training data is not always available. In this publication Genetic Algorithms are used to optimise training data sets. The approach is tested with an existing hydrological model in The Netherlands. The optimised training set resulted in significant better training data.

1 Introduction

Artificial Neural Networks are powerful tools to simulate complex processes under the condition that input and output data sets are available. In hydrology Artificial Neural Networks (ANNs) prove to be good alternatives for traditional modelling approaches. This is particular the case for rainfall-runoff modelling (Minns and Hall, 1996; Whigham and Crapper, 2001; Vos and Rientjes, 2005) and hydraulic flow modelling. The structure of ANNs consists of neurons positioned in layers that are connected through weights and transfer functions. It is in general not straightforward to design a good structure for an ANN; a few rules of thumb for ANN in hydrinformatics were found by Zijderveld (2003). In the training phase the exact values for weights of the network are determined by using one of the available training algorithms, for example the Levenberg-Marquardt backpropagation training function. In this training phase the model actually learns the behaviour of the process by adopting the input-output relations from the data sets. There are several good descriptions on ANN (Hagan et al., 1996; Haykin, 1999). The data set is usually divided into a train and test set and optionally in a cross-validation set. The training data should be representative for the simulation period, otherwise extrapolation of model simulation is difficult. In a flow model for example, the training data must contain high and low flows and in a rainfall-runoff model the data should contain sufficient extreme rainfall events to be representative. Such data is not always available. Considerations for expenditures on sensors, installation, calibration and validation of the data play a role. Data is also limited due to legal, social and technical constraints on its collection and distribution (Shannon et al., 2005). Lack of data is especially the problem for physically distributed hydrological models (Feyen et al., 2000). Using existing data sets may seem a good alternative. However, locations not always match with the problem area and additional problems can occur on the data quality, validation and format. In practise there can also be legal and strategic aspects that give problems to obtain enough validated data or poor data with noise (Doan et al., 2005). In this publication the technique of Genetic Algorithms (GA) is used to reduce this problem by optimising training data sets.

2 Genetic Algorithms

Genetic Algorithms are one of the most successful optimisation techniques of a new generation of soft computing which includes fuzzy-logic, ANN and support-vector machines. From biological sciences, evolutionary processes have been translated to efficient search and design strategies. Genetic Algorithms use these strategies to find an optimum solution for any multi-dimensional problem (Goldberg, 1989). GAs work with children (off-spring) where each child is a copy of its parents plus a variation. After every model run (epoch) the GA creates a new set of children. In this paper the GA is used to feed an ANN model which mimics a hydrological model. Each child consists of a set of five starting points corresponding with the five sub-sets of data of the training data. After several runs, the GA optimised the starting points and therefore the training data of the ANN. The GA solves this problem in a reasonable time without restricting itself to local minima. In this paper the GA constructs the most efficient training data for an ANN network of a hydraulic flow model. From the available dataset usually a subset for training is selected without a predefined selection procedure. The question is whether it is possible to optimise the training data using a GA. The optimisation consists of a procedure which selects area's in the original data set that have a positive influence on the ANN's performance. The GA will search in an evolutionary way if some parts of the data set contain more relevant system knowledge than other parts.

3 The case

In this paper the ANN simulates an existing hydraulic model. It is a model built in Duflow for 1-D free surface flow applied to the drainage system of Baambrugge, The Netherlands by Witteveen+Bos, Fig. 1. The datasets consist of simulated discharges up and downstream of a channel section and is simulated in a one dimensional hydraulic network. The hydraulic network uses the Saint-Vernant equations to solve the free surface flow movements. ANN can simulate free surface flow (e.g. Bobovic and Abbott, 1997; Price et al., 1998; Bobovic and Keijzer, 2000; Minns, 2000; Dibike, 2002). The model has several discharge points and boundaries; there are also many rainfall-runoff area's connected. The input for the ANN is a discharge at the upper boundary. From that point the water disperses into the model. The output of the ANN is the discharge at a point, in the centre of the model where boundary distortion is minimal. The initial data set was constructed in a way that all discharge levels and discharge variations occurred within the limits of the model's hydraulic constrictions. ANNs are capable to generalise system behaviour from the training data (Anctil and Lauzon, 2004) and in a certain extent extrapolate (Shrestha et al., 2005). From this initial training data the GA randomly selects five new sub-sets; all with the same length. The GA selects the starting point of a subset and puts the five sub-sets together in a new training data. This results in a copy of the initial training data, with data standing in a different sequence. Subsequently the ANN performs a run; the outcome is compared with the target values from the test set and the root mean squared-error (RMSE) is calculated. The RMSE is a common measure for the ANN's performance. On the basis of the performance, the GA constructs a new set of starting points using selections, mutations, crossovers and other evolutionary methods. The expectation is that the GA constructs an optimised training set with higher performance than the initial dataset.

4 ANN model experiment

The input for the ANN is the upstream discharge boundary of the model which lies between 0.5 and 1.5 m^3/s. The discharge frequency is slowly increasing from days to hours. The output or training target is the discharge in the central area of the model. The target values are calculated in the numerical flow model. A stepped delay line is used to simulate flow dynamics. In a stepped delay line the input at time t until n steps in history x_{t-n} form the ANN's input layer (see Fig. 2). The target values are the flow models output at time t. The ANN has an input and output layer and also one hidden layer. The network structure, training algorithms, neuron functions and other ANN design parameters are chosen on the basis of other hydrological ANN model designs and experiences. Three differently shaped test sets were constructed.

In this publication the GA is used to optimise the training data of an ANN. The GA is trained for 30 generations each with 10 off-spring. From the initial training set five sub-sets of equal length are selected. The starting point or starting index is chosen at random by the GA (see Fig. 3). The five sub-sets together form a new training set with new input/output time series.

With the combined training set, the ANN makes a calculation, resulting in a new model performance (RMSE). Based on these results the GA starts the next run by choosing five new sample indices.

5 Conclusions

For many reasons sufficient and representative data is not always available. In this case we focused on a hydraulic flow model and presumed there is not enough data to train an ANN. Therefore a dataset was composed regarding the properties of the flow network such as water depth, average discharge and timescales. With this dataset the ANN simulated and predicted model output which gave, as expected, poor results.

The GA was used to improve the results of the ANN by optimising the original training set. The algorithm selected five sub-sets from that training set and placed them in a new data set. This was repeated in an optimised algorithm with a evolutionary growth background that resulted in a training set that performed much better. In this particular flow model it gave on average 39% better results when measured in RMSE. Figure 4 shows one of the resulting training sets. The sharp edges in the discharge indicate the borders of the sub-sets and have no special meaning.

To take a closer look to the resulting data set, Fig. 5 shows the distribution of training samples of the five best performing and optimised training sets in comparison with the original samples. A sample with a high density indicates that a sample data is often selected in all sub-sets.

From this figure it is shown that there are no special peaks or trends. The GA used data from the entire, original dataset consisting of 1489 data samples to find an optimised training set, except for the first 77 and the last 372 samples. In the last area the frequencies of the discharge is very high as shown in Fig. 3. The explanation for this is that in the flow model quick changes result in an almost constant discharge in the centre of the model. The effect is that the relation between inputs and outputs is not ambiguous anymore. An ANN cannot handle this. As a result the GA did not select this area in the optimised training set. Slower changes were no problem for the ANN except for the very first samples where noise, induced by initial conditions, influenced the results.

In this paper a GA was used to optimise the training data for an ANN simulating an existing hydrological flow model in Baambrugge, The Netherlands. The resulting training data was built from five sub-sets selected by the GAs optimisation technique and resulted in an ANN which gives more accurate outputs.

References

Anctil, F. and Lauzon, N.: Generalisation for neural networks through data sampling and training procedures with applications to streamflow predictions, Hydrol. Earth Syst. Sci., 8, 940–958, 2004. 288

Babovic, V. and Keijzer, M.: Genetic programming as a model induction engine, J. Hydroinformatics, 2, 1, 35–60, 2000. 288

Babovic, V. and Abbott, M. B.: The evolution of equations from hydraulic data: Part I – Theory, J. Hydraulic Res., 35, 3, 1–14, 1997. 288

Dibike, Y. B.: Model Induction from Data: Towards the next generation of computational engines in hydraulics and hydrology, IHE Delft, Delft, 2002. 288

Doan, C. D., Liong, S. Y., and Karunasinghe, D. S. K.: Derivation of effective and efficient data set with subtractive clustering method and genetic algorithm, J. Hydroinformatics, 7, 219–233. 287

Feyen, L., Vázquez, R., Christiaens, K., Sels, O., and Feyen, J.: Application of a distributed physically-based hydrological model to a medium size catchment, Hydrol. Earth Syst. Sci., 4, 47–63, 2000. 287

Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Pub. Co., 1989. 287

Hagan, T., Demuth, H. B., and Beale, M. H.: Neural Network Design, PWS Pub. Co., Boston, 1996. 286

Haykin, S.: Neural Networks, a comprehensive foundation, Prentice Hall, New Jersey, 1999. 286

Minns, A. W. and Hall, M. J.: Artificial neural networks as rainfall-runoff models, Hydrol. Sci. J., 41, 3, 399–417, 1996. 286

Minns, A. W.: Subsymbolic methods for data mining in hydraulic engineering, J. Hydroinformatics, 2, 3–13, 2000. 288

Price, R. K., Samedov, J. N., and Solomatine, D. P.: An artificial neural network model of a generalised channel network, Proc. 3rd Int. conf. Hydroinformatics, Copenhagen, 1998. 288

Shannon, C., Moore, D., Keys, K., Fomenkov, M., Huffaker, B., and Claffy, K.: The internet measurement data catalog, Computer communication review, 35, 97–100, 2005. 287

Shrestha, R. G., Theobald, S., and Nestmann, F.: Simulation of flood flow in a river system using artificial neural networks, Hydrol. Earth Syst. Sci., 9, 313–321, 2005. 288

de Vos, N. J. and Rientjes, T. H. M.: Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation, Hydrol. Earth Syst. Sci., 9, 111–126, 2005, 286

Whigham, P. A. and Crapper, P. F.: Modelling rainfall-runoff using genetic programming, Mathematical and Computer Modelling, 33, 6–7, 707–721, 2001. 286

Zijderveld, A.: Neural network design strategies and modelling in hydroinformatics, Ph.D. thesis, Delft University of Technology, Delft, 2003. 286

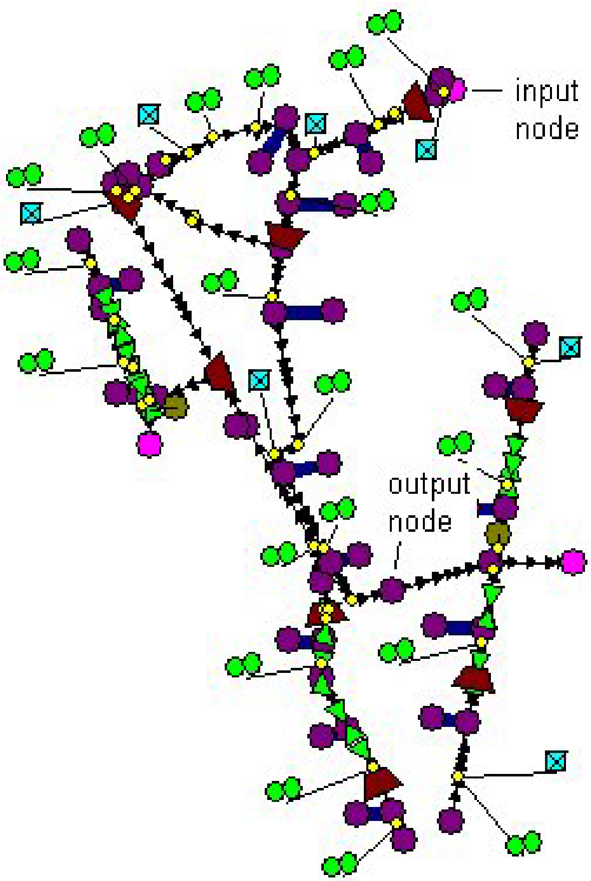

Fig. 1. Hydrological flow model Baambrugge, the Netherlands (Witteveen+Bos).

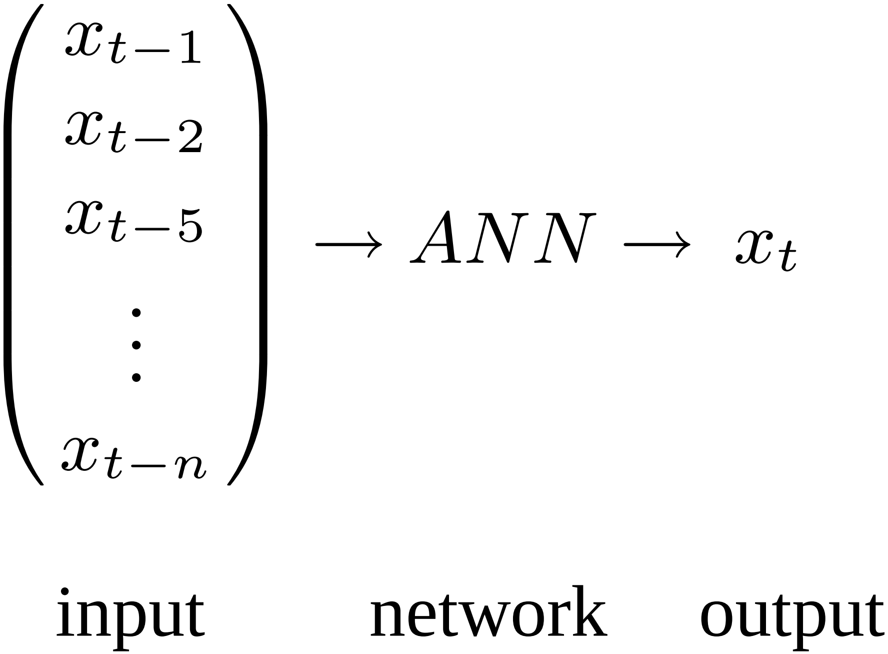

Fig. 2. Input with stepped delay line to simulate history.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◀ | ▶

◀ | ▶

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

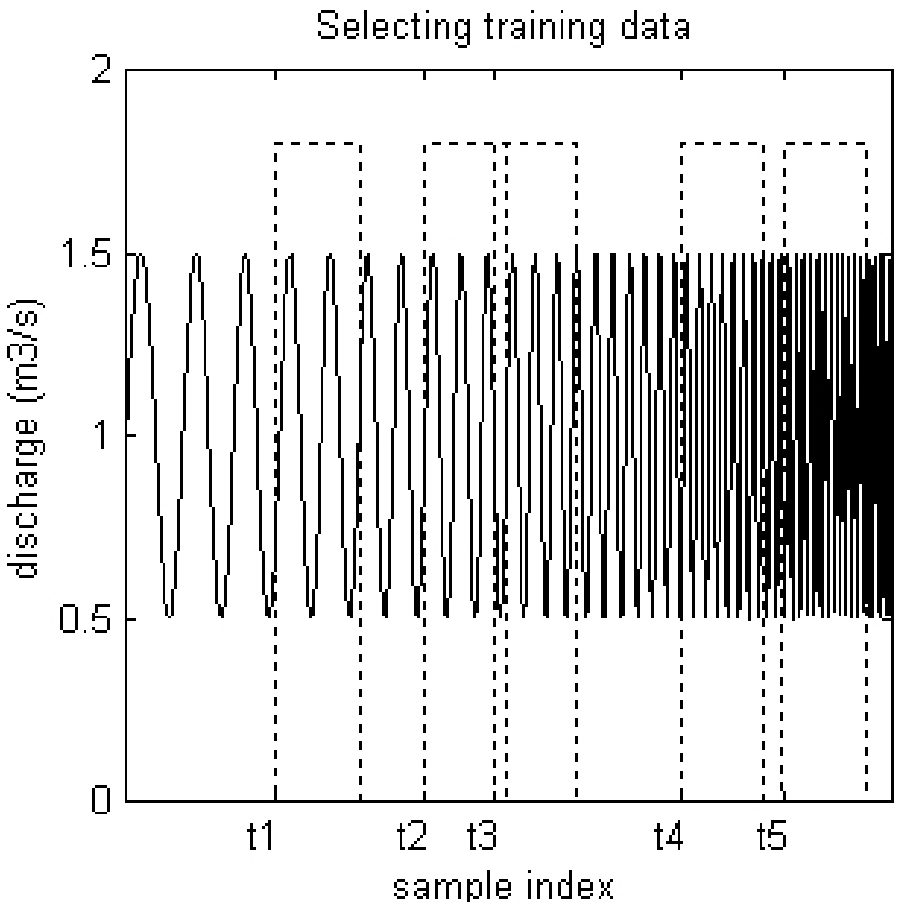

Fig. 3. Example of starting index for five sub-sets in training data.

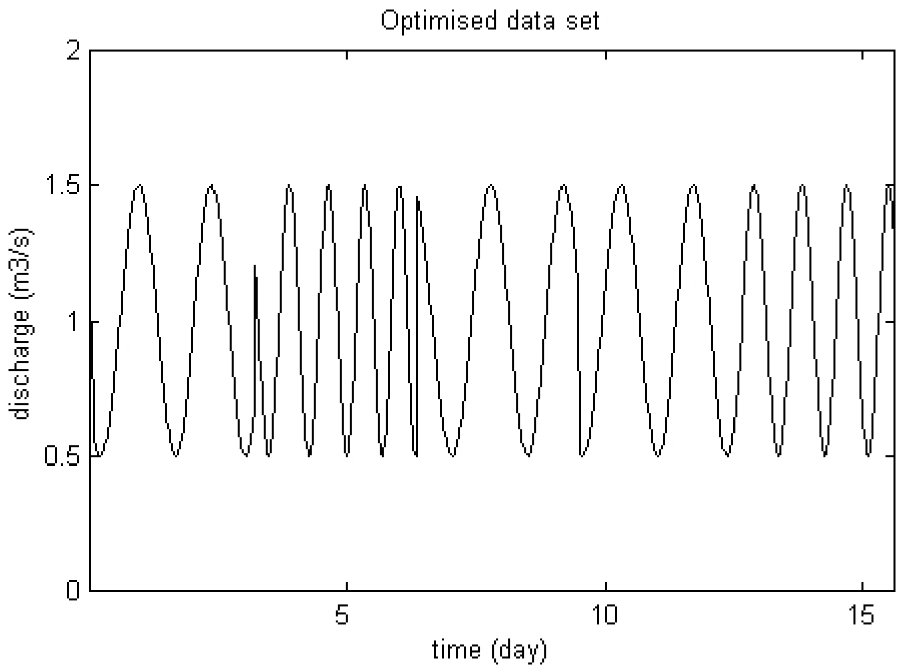

Fig. 4. Training set optimised by Genetic Algorithm.

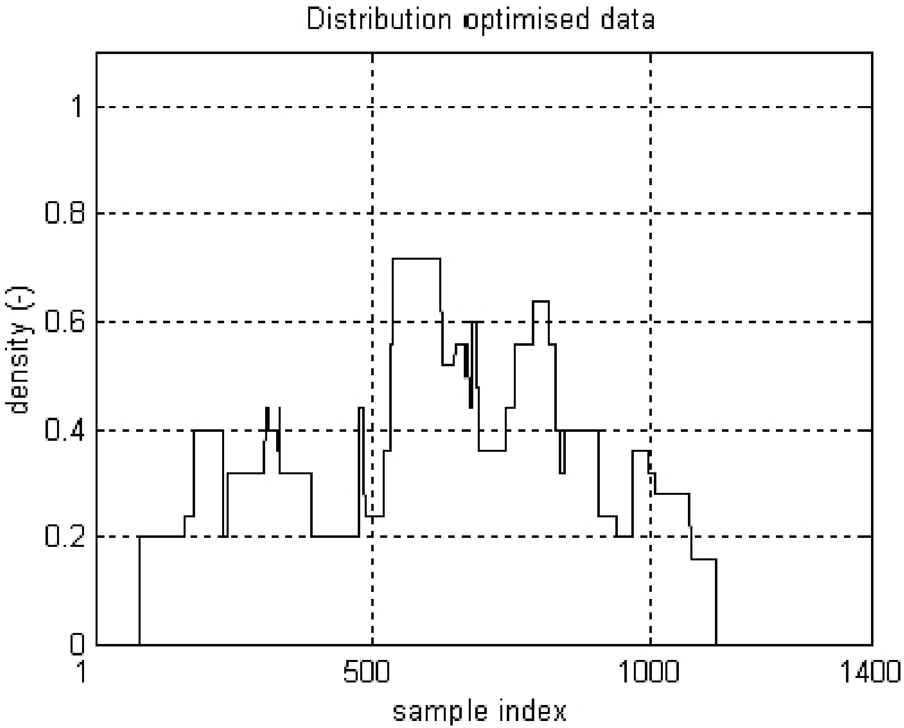

Fig. 5. Distribution sub-set data in optimised training set.