ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 407-412

Neural networks and M5 model trees in modeling water level-discharge relationship for an Indian river

B. Bhattacharya

D.P. Solomatine

International Institute for Infrastructural, Hydraulic, and Environmental Engineering
(IHE), Delft, The Netherlands
E-mail: *{bha, sol}@ihe.nl*

Abstract: In flood management it is important to reliably estimate the discharge in a river. Hydrologists use historic data to establish a rating curve – a relationship between the water level (stage) and discharge. ANN and M5 model trees were used to reconstruct this relationship on an example of an Indian river. The predictive accuracy of these machine learning methods models was found to be superior to a conventional rating curve.

1. Introduction

In flood management it is important to reliably estimate the discharge in a river. A functional relationship between the water level (also called the stage) and discharge is established with the help of field measurements and the relationship is expressed as a rating curve. Normally a polynomial regression equation is used to represent a rating curve, or regression- and auto-correlation-based statistical methods such as ARIMA models can be used. However, the use of function approximation methods related to machine learning could be a better alternative.

ANN is the most widely accepted machine learning method and is widely used in various areas of water-related research such as rainfall-runoff modelling (Dawson and Wilby 1998; Dibike and Solomatine 2000), prediction of discharge (Muttiah et al., 1997). ANNs were found to be very efficient in modelling stage-discharge relationship (Bhattacharya and Solomatine, 2000; Jain and Chalisgaonkar, 2000). Such machine learning technique as a M5 model tree (MT, Quinlan 1992) is less known but it is a promising numerical prediction method that has been proved to be very efficient and robust. MT is not yet as popular as ANN, and, for example in the water sector its use started only recently (Kompare 1997; Solomatine and Dulal, 2003).

In the present paper ANN and MT models of the stage-discharge relationship at one discharge measuring station have been compared with a conventional rating curve. MLP ANN, as a widely accepted method will not be presented here; rather more space will be given to model trees.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 407-412

2. Model tree: an introduction

One of a popular ways of classification (where the task consists of assigning a particular input example, or vector, to a class) is a decision tree (DT). DT consists of leaf or answer nodes that indicate a class and non-leaf or decision nodes that contain an attribute name and branches to other decision trees, one for each value of the attribute. The top-down induction of decision trees is a popular approach in which classification starts from a root node and proceeds to generate sub-trees until leaf nodes are created. There are several efficient algorithms for building decision trees such as ID3 and C4.5 (Quinlan, 1986).

In regression (numerical prediction) problems DTs cannot be applied. However, the success with decision trees in the classification problems has motivated researchers to extend this method to the regression problems by introducing ranges in the numeric values of the output so that it can be treated as a class. One of such methods is a *regression trees* where the leaf nodes contain a constant numeric value (that is a zero*th* order regression model) which is the average of all the training set values that the leaf applies to (Breiman et al, 1984).

There are two other methods able to generate more complex, 1st order (linear) models: the approach by Friedman (1991) in his MARS (*multiple adaptive regression splines*) algorithm, and the one used in this paper, *M5 model tree* (Quinlan 1992; Witten and Frank, 2000).

The structure of MT follows that of decision trees and has multivariate linear regression models at the leaf nodes. Thus an MT is a combination of piecewise linear models each of which is suitable for a particular domain of input space (Fig. 1). The algorithm of an MT breaks the input space of the training data through nodes or decision points to assign a linear model suitable for that sub-area of the input space. The continuous splitting often results in a too complex tree that needs to be pruned (reduced) to a simpler tree to improve the generalisation capacity. Finally, the value predicted by the model at the appropriate leaf is adjusted by the smoothing operation to reflect the predicted values at the nodes along the path from the root to that leaf.

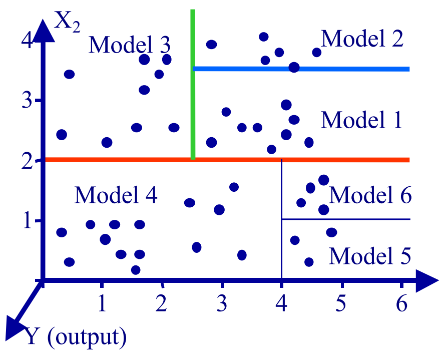

Fig. 1 Splitting the input space X_1 x X_2 by M5 model tree algorithm; each model is a linear regression model $y = a_0 + a_1x_1 + a_2x_2$

The overall global model, which is the collection of these linear (and locally accurate) models brings the required non-linearity in dealing with the problem. The difference from pure linear regression is that the necessary (sub)optimal splitting of input space is performed automatically. MTs can learn efficiently and can tackle tasks with high dimensionality which can be up to hundreds of attributes. The resulting MTs are transparent and simple – this makes them potentially more successful in the eyes of decision makers.

4. Experimental set up

A widely used conventional relationship for the rating curve is expressed as $Q = \alpha(h\text{-}h_0)^{\beta}$ (where h_0 stands for the minimum stage below which a discharge is not feasible, h is stage and Q is discharge), and the values for α and β are chosen so that they maximise the fit to the training data. We used the rating curve that has already been used in practice and calibrated, see Fig. 2. (Note that the identification of the rating curve is in fact also a function approximation problem, same as solved by an ANN; the idea of the experiment was to test how well the other methods work.)

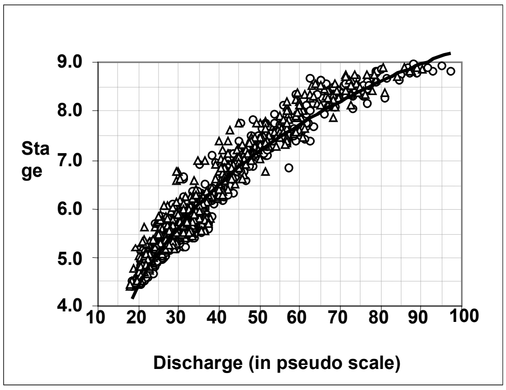

Fig. 2 Conventional stage-discharge model at Swarupgunj

Data for the period 1990 to 1998 was from a discharge measuring station at Swarupgunj on the river Bhagirathi in India has been considered. It is unidirectional with a width of about 320 m and maximum depth of about 8 m. Number of training and verification examples were 1364 and 621 respectively. The model had to predict the discharge Q at the next time step $t+1$ by reconstructing the relationship $Q_{t+1} = f(h_{t+1}, h_t, h_{t-1}, Q_t)$.

For building MT the *Weka* software was used (Witten and Frank, 2000). The ANN model was built with *NeuralMachine* (*www.data-machine.com*). We used a MLP ANN with the backprop training, one hidden layer, logistic transfer functions; the number of hidden nodes was optimised. We used PC with the Pentium 3 at 600MHz. Training of ANN took 10 minutes and of MT only 4 sec. Execution time on verification data set is negligible (less than 0.5 sec for both models). Development of each model took two to three weeks.

Table 1 Comparison of errors in MTs of different complexity
(RMSE=root mean squared error; NRMSE=normalized RMSE)

Number of linear models	Training		Verification	
	RMSE	NRMSE	RMSE	NRMSE
94	79.3	0.132	76.0	0.111
4	89.8	0.150	69.1	0.101
2	92.0	0.153	69.7	0.101

4. Results and discussions

The first MT generated was very complex with 94 linear models at the leaf nodes. It was very accurate in training but overfit and had to be pruned in order to ensure good generalisation capacity. Pruning is done until the predictive accuracy does not drop substantially. Table-1 shows the performance of the three model versions. The model with 4 leaves (linear models) is given below:

if Q_t <= 37.5 then
  if Q_t <= 28.25 then Q_{t+1} = -243 – 187 h_t-1 + 299 h_t + 0.667 Q_t
  if Qt > 28.25 then Q_{t+1} = -214 – 387 h_t-1 + 448 h_t + 0.885 Q_t
if Q_t > 37.5 then
  if h_t <= 7.85 then Q_{t+1} = -455 – 491 h_{t-1} + 628 h_t + 0.727 Q_t
  if h_t > 7.85 then Q_{t+1} = -1720 – 605 h_{t-1} + 924 h_t + 0.66 Q_t

From Table-1 it can be seen that without loosing too much accuracy a model with only 2 linear models can be adopted; its equations are as follows:

if $Q_t \le 37.5$ then Q_{t+1} = –204-301 h_t + 383 h_{t-1} + 0.788 Q_t
if $Q_t > 37.5$ then Q_{t+1} = –728-550 h_t + 721 h_{t-1} + 0.745 Q_t

It is interesting to note that from this pruned model the term h_{t-1} has disappeared though it was present in more complex models. The discharge hydrograph predicted by this MT along with the measured discharge hydrograph is plotted in Fig. 3.

Table 2 Performance and training times for different models

	Training		Verification		Duration of training, *s*
	RMSE	NRMSE	RMSE	NRMSE	
Model tree	92.0	0.153	69.7	0.101	4
ANN	90.5	0.151	70.5	0.103	1200
Conventional rating curve	143.3	0.239	111.2	0.162	n/a

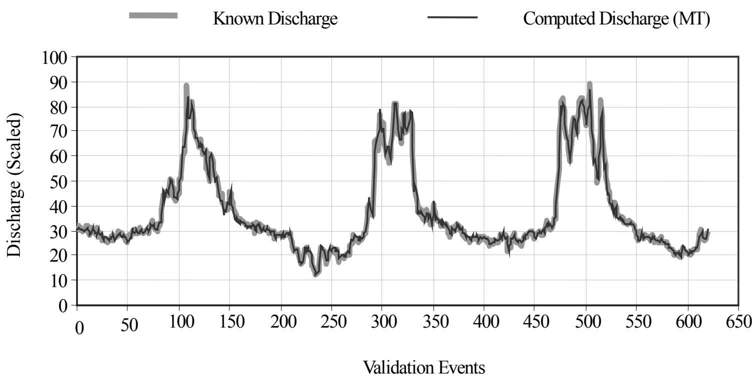

Fig. 3 Discharge predicted by the MT vs with the known discharge (the ANN-generated plots are very similar)

Training and testing errors of MT and ANN models are very close to each other (Table-2). Both machine learning models have out-performed the conventional rating curve.

5. Conclusions

Since rating curve development is associated with the collection of considerable amount of data, the use of machine learning methods appeared to be justified. The predictive accuracy of the simplest MT model was observed to be very high and at par with that of an ANN model built with the same data. The advantage of MT appeared to be in being transparent, giving an expert a very simple and easily verifiable model. Both ANN and MT were found to be considerably better than the conventional rating curve.

Part of this work is part of the project "Data mining, and data-driven modelling" (Delft Cluster programme) supported by the Dutch government.

References

Bhattacharya, B. and Solomatine, D.P. (2000). “Application of artificial neural network in stage discharge relationship”, *Proc. of 4th Int. Conf. on Hydroinformatics* , Iowa, USA.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees,* Wadsworth, CA.

Dawson C. W. and Wilby R. (1998). “An artificial neural network approach to rainfall-runoff modelling”, *Hydrological Sci. J.*, **43** (1), 47-66.

Dibike Y.B. and Solomatine D.P. River Flow Forecasting Using ANNs, Journal of Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere, 2001, **26** (1), 1-8.

Friedman, J.H. (1991). “Multivariate adaptive regression splines”, *Annals of Statistics*, **19**, 1-141.

Jain, S.K. and Chalisgaonkar, D. (2000). “Setting up stage-discharge relations using ANN”, *J. of Hydrologic Engg.*, ASCE, **5 (**4), 428-433.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 407-412

Kompare, B., Steinman, F., Cerar, U., and Dzeroski, S. (1997). Prediction of rainfall runoff from catchment by intelligent data analysis with machine learning tools within the artificial intelligence tools. *Acta hydrotechnica* (in Slovene language) 16/16.

Muttiah R.S., Srinivasan R., Allen P.M (1997). “Prediction of two-year peak stream discharges using neural networks”, *J. of Am. Water Res. Assoc.*, **33** (3), 625-630.

NeuralMachine (2003). *www.data-machine.com.*

Quinlan, J.R. (1986). “Induction of decision trees”, *Machine learning*, Vol.1, 181-186.

Quinlan, J.R. (1992). “Learning with continuous classes”, *Proc. of Australian Joint Conf. on AI*, 343-348, World Scientific, Singapore.

Solomatine, D.P. and Dulal, K.N. (2003). "Model tree as an alternative to neural network in rainfall-runoff modelling". *Hydrological Sciences J. (to appear).*

Westphal, J.A., Thompson, D.B., Stevens, G.T., Strauser, C.N. (1999). “Stage-discharge relations on the middle Mississippi river”, *J. of Wat. Res. Plng. & Mgmt.*, ASCE, **125**, No. 1, 48-53.

Witten, I.H. and Frank, E. (2000). *Data mining*, Morgan Kaufmann.