Geophysical Research Abstracts, Vol. 9, 01391, 2007
SRef-ID: 1607-7962/gra/EGU2007-A-01391

Applying an instance-based learning approach to the Bird Creek dataset

L.M. See (1), A.J. Heppenstall (1)

(1) School of Geography, University of Leeds, Leeds, LS2 9JT, UK, (l.m.see@leeds.ac.uk / Fax: +44 113-3433308)

Instance-based learning involves combining instances from a data set that are close in attribute space to the input vector for which a prediction is required. The algorithm is derived from a nearest neighbour classifier and can be generalised to a $k-$nearest neighbour method, where the value of k can be specified by the user or determined through cross-validation. A simple average is then used to combine the instances or alternatively, a weighted distance decay function can be applied. Instance-based learning was applied to the Bird Creek calibration dataset as part of the EGU Hydrological Forecasting Competition (HS47) using the Weka data mining software. The best performing model on the calibration data set for a lead time of 6 hours had the following inputs: Flow at time t, Flow at t-6 hours and Rainfall at t-18 hours. For a lead time of 24 hours, the following inputs produced the best performing model: Flow at time t, t-6 and t-12 hours and Rainfall at t-18, t-24, t-30 and t-36 hours. Ten-fold cross-validation was used to determine the optimal number of nearest neighbours, which was 4 for a lead time of 6 hours and 7 for a lead time of 24 hours. Both models also incorporated inverse distance weighting. The performance measures for the model on the independent test dataset for a lead time of 6 hours were an r-squared of 0.9560, a root mean squared error (RMSE) of 29.76 and a mean absolute error (MAE) of 8.66. For a lead time of 24 hours the r-squared was 0.3457, with a RMSE of 109.62 and a MAE of 39.72.