Geophysical Research Abstracts, Vol. 9, 09855, 2007
SRef-ID: 1607-7962/gra/EGU2007-A-09855

M5 model tree applied to the HFC Bird Creek Data Set

R.J. Abrahart (1) and L.M. See (2)

(1) School of Geography, University of Nottingham, UK; (2) School of Geography, University of Leeds, UK (Email: bob.abrahart@nottingham.ac.uk)

Weka 3.4 (Witten & Frank, 2005; http://www.cs.waikato.ac.nz/˜ml/weka/) is the leading open-source project in machine learning. It comprises a comprehensive collection of machine-learning algorithms for data mining tasks. This software package contains tools for data pre-processing, classification, regression, clustering, visualization, etc. It also provides a popular tool for the development of decision tree models. There are two main types of decision trees: [1] classification trees are the most common type and are used to predict a symbolic attribute, which is called the class; and [2] regression trees which are be used to predict the value of a numeric attribute. If each leaf in the tree contains a linear regression model, that is used to predict the target value at that leaf, it is called a model tree. Thus regression trees and model trees are similar since at each leaf in the tree a numeric output rather than categorical output is produced. The two nevertheless differ in the exact nature of this output; regression trees produce an averaged numeric prediction for each leaf in the tree, whereas model trees, that have a linear equation at each leaf, will produce an exact prediction on each occasion. Model trees have other advantages over regression trees in terms of compactness and prediction skill which arises out of their power to exploit local linearities in the dataset. Model trees can also extrapolate. Moreover, by dividing the function that is being introduced into linear patches, model trees will provide a representation that is both reproducible and somewhat more comprehensible in comparison to less transparent solutions such as neural networks

This paper will present the results of an M5 model tree that was applied to the contest dataset and developed using the default parameter settings.

Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco