A branch and bound algorithm for computing the best subset regression models

Cristian Gatu and Erricos John Kontoghiorghes
Institut d'informatique, Université de Neuchâtel,
Emile-Argand 11, CP 2, CH-2007 Neuchâtel, Switzerland
{cristian.gatu, erricos.kontoghiorghes}@unine.ch

Keywords: Subset regression; Least squares; Branch and bound algorithm; QR decomposition.

Abstract

An efficient branch and bound algorithm for computing the best subset regression models is proposed. The algorithm avoids the computation of the whole regression tree that generates all possible subset models. The criterion used in identifying the best subsets are based on the residual sum of squares (RSS). Various statistics associated with model selection are function of RSS. The algorithm has the QR decomposition as a main computational tool and outperforms an existing leaps and bounds strategy based on inverse matrices. Strategies and heuristics which improve the computational performance of the proposed algorithm are investigated. A computational efficient heuristic version of the branch and bound algorithm which decides to cut subtrees using a tolerance parameter is proposed. The heuristic algorithm derives models closed to the best ones. However, it is shown analytically that the relative error of the RSS of the computed subsets is smaller than the value of the tolerance parameter which lies between zero and one. The computational speed of the proposed algorithms is analyzed using theoretical measures of complexity and simulations on random data sets. The performance of the selection strategies is evaluated using real experimental data.