Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 16–29
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Efficient algorithms for computing the best subset regression models for large-scale problems☆

Marc Hofmann[a,*], Cristian Gatu[a,d], Erricos John Kontoghiorghes[b,c]

[a]*Institut d'Informatique, Université de Neuchâtel, Switzerland*
[b]*Department of Public and Business Administration, University of Cyprus, Cyprus*
[c]*School of Computer Science and Information Systems, Birkbeck College, University of London, UK*
[d]*Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi, Romania*

Available online 24 March 2007

Abstract

Several strategies for computing the best subset regression models are proposed. Some of the algorithms are modified versions of existing regression-tree methods, while others are new. The first algorithm selects the best subset models within a given size range. It uses a reduced search space and is found to outperform computationally the existing branch-and-bound algorithm. The properties and computational aspects of the proposed algorithm are discussed in detail. The second new algorithm preorders the variables inside the regression tree. A radius is defined in order to measure the distance of a node from the root of the tree. The algorithm applies the preordering to all nodes which have a smaller distance than a certain radius that is given a priori. An efficient method of preordering the variables is employed. The experimental results indicate that the algorithm performs best when preordering is employed on a radius of between one quarter and one third of the number of variables. The algorithm has been applied with such a radius to tackle large-scale subset-selection problems that are considered to be computationally infeasible by conventional exhaustive-selection methods. A class of new heuristic strategies is also proposed. The most important of these is one that assigns a different tolerance value to each subset model size. This strategy with different kind of tolerances is equivalent to all exhaustive and heuristic subset-selection strategies. In addition the strategy can be used to investigate submodels having noncontiguous size ranges. Its implementation provides a flexible tool for tackling large scale models.

Keywords: Best-subset regression; Regression tree; Branch-and-bound algorithm

1. Introduction

The problem of computing the best-subset regression models arises in statistical model selection. Most of the criteria used to evaluate the subset models rely upon the residual sum of squares (RSS) (Searle, 1971; Sen and Srivastava, 1990). Consider the standard regression model

$$y = A\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_m), \tag{1}$$

☆ The R routines can be found at URL: ⟨http://iiun.unine.ch/matrix/software⟩.

* Corresponding author. Tel.: +41 32 7182708; fax: +41 32 7182701.
E-mail addresses: marc.hofmann@unine.ch (M. Hofmann), cristian.gatu@unine.ch (C. Gatu), erricos@ucy.ac.cy (E.J. Kontoghiorghes).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.017

Table 1
Leaps and BBA: execution times in seconds for data sets of different sizes, without and with variable preordering

# Variables	36	37	38	39	40	41	42	43	44	45	46	47	48
Leaps	8	29	44	30	203	57	108	319	135	316	685	2697	6023
BBA	2	5	12	8	35	14	9	55	27	37	97	380	1722
Leaps-1	3	16	28	9	82	33	22	203	79	86	306	1326	1910
BBA-1	1	4	13	2	20	11	4	47	18	15	51	216	529

where $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ is the exogenous data matrix of full column rank, $\beta \in \mathbb{R}^n$ is the coefficient vector and $\varepsilon \in \mathbb{R}^n$ is the noise vector. The columns of A correspond to the exogenous variables $V = [v_1, \ldots, v_n]$. A submodel S of (1) comprises some of the variables in V. There are $2^n - 1$ possible subset models, and their computation is only feasible for small values of n. The dropping column algorithm (DCA) derives all submodels by generating a regression tree (Clarke, 1981; Gatu and Kontoghiorghes, 2003; Smith and Bremner, 1989). The parallelization of the DCA moderately improves its practical value (Gatu and Kontoghiorghes, 2003). Various procedures such as the forward, backward and stepwise selection try to identify a subset by inspecting very few combinations of variables. However, these methods rarely succeed in finding the best submodel (Hocking, 1976; Seber, 1977). Other approaches for subset-selection include ridge regression, the nonnegative garrote and the lasso (Breiman, 1995; Fan and Li, 2001; Tibshirani, 1996). Sequential replacement algorithms are fairly fast and can be used to give some indication of the maximum size of the subsets that are likely to be of interest (Hastie et al., 2001). The branch-and-bound algorithms for choosing a subset of k features from a given larger set of size n have also been investigated within the context of feature selection problems (Narendra and Fukunaga, 1997; Roberts, 1984; Somol et al., 2004). These strategies are used when the size k of the subset to be selected is known. Thus, they search over $n!/(k!(n-k)!)$ subsets.

A computationally efficient branch-and-bound algorithm (BBA) has been devised (Gatu and Kontoghiorghes, 2006; Gatu et al., 2007). The BBA avoids the computation of the whole regression tree and it derives the best subset model for each number of variables. That is, it computes

$$\underset{S}{\operatorname{argmin}}\ \mathrm{RSS}(S) \quad \text{subject to } |S| = k \ \text{ for } k = 1, \ldots, n. \tag{2}$$

The BBA was built around the fundamental property

$$\mathrm{RSS}(S_1) \geqslant \mathrm{RSS}(S_2) \quad \text{if } S_1 \subseteq S_2, \tag{3}$$

where S_1 and S_2 are two variable subsets of V (Gatu and Kontoghiorghes, 2006). The BBA-1, which is an extension of the BBA, preorders the n variables according to their strength in the root node. The variables i and j are arranged such that $\mathrm{RSS}(V_{-i}) \geqslant \mathrm{RSS}(V_{-j})$ for each $i \leqslant j$, where V_{-i} is the set V from which the ith variable has been deleted. The BBA-1 has been shown to outperform the previously introduced leaps-and-bounds algorithm (Furnival and Wilson, 1974). Table 1 shows the execution times of the BBA and leaps-and-bounds algorithm for data sets with 36–48 variables. Note that the BBA outperforms the leaps-and-bounds with preordering in the root node (Leaps-1). A heuristic version of the BBA (HBBA) that uses a tolerance parameter to relax the BBA pruning test has been discussed. The HBBA might not provide the optimal solution, but the relative residual error (RRE) of the computed solution is smaller than the tolerance employed.

Often models within a given size range must be investigated. These models, hereafter called subrange subset models, do not require the generation of the whole tree. Thus, the adaptation of the BBA for deriving the subrange subset models is expected to have a lower computational cost, and thus, it can be feasible to tackle larger scale models. The structural properties of a regression tree strategy which generates the subrange subset models is investigated and its theoretical complexity derived. A new nontrivial preordering strategy that outperforms the BBA-1 is designed and analyzed. The new strategy, which can be found to be significantly faster than existing ones, can derive the best subset models from a larger pool of variables. In addition, some new heuristic strategies based on the HBBA are developed. The tolerance parameter is either a function of the level in the regression tree, or of the size of the subset model. The novel strategies decrease execution time while selecting models of similar, or of even better, quality.