ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
21,3 (1992), pp. 343–347

A. BARTKOWIAK (Wrocław)

FINDING AN ε-OPTIMAL REGRESSION SUBSET

1. Recall of a branch and bound algorithm. In [1] we presented an algorithm for finding an optimal subset of size k out of p predictor variables $x_1, \ldots, x_p$ by use of a branch and bound method. Instead of considering all $\binom{p}{k}$ subsets to find the best one for predicting a variable y we proposed an algorithm which proceeds stepwise in the following manner:

All possible $\binom{p}{k}$ subsets are divided into parts called branches. For each branch a bound B_i $(i = 1, \ldots, k)$ for RSS_i, the residual sum of squares (a criterion of "prediction goodness" of a subset $x_{i_1}, \ldots, x_{i_k}$) is established. No subset belonging to the ith branch can yield a smaller value of RSS than the established bound B_i.

The bounds $B_1, \ldots, B_k$ for branches $1, \ldots, k$ satisfy the inequalities

$$B_1 \geq B_2 \geq \ldots \geq B_k .$$

For $i = k,\ k-1, \ldots, 1$ we investigate all subsets belonging to the ith branch. Suppose that after investigation of branches $k,\ k-1, \ldots, i,\ i > 1$, we found the optimal subset among those branches to be $\{x^0_{i_1}, \ldots, x^0_{i_k}\}$, with residual sum of squares RSS^0 satisfying the inequality

$$\mathrm{RSS}^0 \leq B_{i-1} .$$

Then we conclude that in branches $i-1, \ldots, 1$ there is no better subset, i.e. yielding a smaller RSS.

2. The notion of an ε-optimal subset. Let $P = \{S_1, \ldots, S_l\}$ be the set of all $l = \binom{p}{k}$ combinations $\{x_{i_1}, \ldots, x_{i_k}\}$ of length k constructed from the predictor variables $x_1, \ldots, x_p$ in a regression problem.

DEFINITION 1. A subset $S_\alpha \in P$ is called an *optimal regression subset* if

$$\mathrm{RSS}^{(\alpha)} \leq \mathrm{RSS}^{(\beta)} \tag{1}$$

1991 *Mathematics Subject Classification*: Primary 62J05.

Key words and phrases: linear regression, multiple regression, residual sum of squares, optimal subset, search of optimal subset by branch and bound.

for any subset $S_\beta \in P$, $S_\beta \neq S_\alpha$, where $\mathrm{RSS}^{(\alpha)}$ and $\mathrm{RSS}^{(\beta)}$ are the residual sums of squares when considering the regressions of the variable y from the variables in S_α and S_β, respectively.

An optimal subset need not be unique.

DEFINITION 2. Let $\varepsilon > 0$ be a given small number. A subset S_γ is called an *ε-optimal regression subset* if its residual sum of squares is within ε of an optimal subset S_α, i.e.

$$0 < \mathrm{RSS}^{(\gamma)} - \mathrm{RSS}^{(\alpha)} \leq \varepsilon . \tag{2}$$

For a given ε there may be several, one or no ε-optimal subsets.

THEOREM. *Suppose that $\varepsilon > 0$ is a given small number. Suppose further that we are seeking for an optimal regression subset (in the sense of* (1)) *using the branch and bound algorithm introduced in Section* 1. *Suppose that we have already examined branches k, $k-1, \ldots, i$, $i > 1$, and that we have found a subset S_δ to be optimal so far, yielding a residual sum of squares* $\mathrm{RSS}^{(\delta)}$.

(a) *If*

$$\mathrm{RSS}^{(\delta)} \leq B_{i-1} , \tag{3}$$

where B_{i-1} is the bound established for branch $i-1$, then S_δ is optimal.

(b) *If*

$$0 < \mathrm{RSS}^{(\delta)} - B_{i-1} \leq \varepsilon \tag{4}$$

then S_δ is either optimal or ε-optimal.

Proof. (a) The set S_δ is the best subset found so far and its residual sum of squares satisfies (3). From the method of allotting the subsets to branches and establishing the bounds it follows that every subset S_β belonging to branches $i-1, \ldots, 1$ yields an RSS greater than or equal to B_{i-1}. Therefore

$$\mathrm{RSS}^{(\delta)} \leq B_{i-1} \leq \mathrm{RSS}^{(\beta)} , \tag{5}$$

and we conclude that S_δ is optimal.

(b) Suppose now that S_δ satisfies (4), i.e.

$$B_{i-1} < \mathrm{RSS}^{(\delta)} \leq B_{i-1} + \varepsilon .$$

Consider the set S of all subsets S_τ in branches $i-1, \ldots, 1$ satisfying

$$B_{i-1} < \mathrm{RSS}^{(\tau)} < \mathrm{RSS}^{(\delta)} .$$

If $S = \emptyset$, i.e. all subsets S_τ in branches $i-1, \ldots, 1$ satisfy $\mathrm{RSS}^{(\tau)} \geq \mathrm{RSS}^{(\delta)}$, then clearly S_δ is optimal.

Otherwise one of the sets in S, say S_α, is optimal. Since

$$B_{i-1} \leq \mathrm{RSS}^{(\alpha)} < \mathrm{RSS}^{(\delta)} ,$$

and, from (4),

$$\mathrm{RSS}^{(\delta)} \le B_{i-1} + \varepsilon\,,$$

we conclude that

$$0 < \mathrm{RSS}^{(\delta)} - \mathrm{RSS}^{(\alpha)} \le \varepsilon\,,$$

i.e. S_δ is ε-optimal.

3. The algorithm. It is a straightforward implementation of the Theorem presented in Section 2. For a fixed value of k (size of the subset) we establish the bounds $B_1 \ge \ldots \ge B_k$ for subsequent branches. Next for $m = k,\ k-1, \ldots, 1$ we perform evaluations of the RSS (residual sum of squares) for subsets belonging to the mth branch. Let $\mathrm{RSS}^{(\delta)}$ be the smallest RSS found after investigating the mth branch. If $\mathrm{RSS}^{(\delta)}$ satisfies (3) or (4) then we have found either the optimal or the ε-optimal subset and we stop our search; otherwise we have to continue the investigations for the next value of m.

The algorithm can be modified by considering instead of RSS (the residual sum of squares) the ratio RSS/SST, with $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$. The ratio RSS/SST is connected with R^2, the square of the multiple correlation coefficient $R^2 = R^2_{y.x_1,\ldots,x_k}$ between the variable y and the predictor variables $x_1, \ldots, x_k$ by the following formula :

$$\mathrm{RSS}/\mathrm{SST} = 1 - R^2\,.$$

In this case the constant ε $(0 \le \varepsilon \le 1)$ can be interpreted in terms of the multiple correlation coefficient R: Our algorithm finds a subset $\{x_{i_1}, \ldots, x_{i_k}\}$ yielding a multiple correlation coefficient $\widetilde{R}$ such that

$$|\widetilde{R}^2 - R^2_{\mathrm{opt}}| \le \varepsilon\,,$$

where R_{opt} is the optimal (maximal) multiple correlation coefficient which can be established for the considered regression with k predictor variables.

4. Simulation examples and practice with the algorithm. We have checked the performance of the algorithm on 60 artificially generated sets of data, the same which were used for checking the performance of the branch and bound algorithm described in [1]. These are data with $p = 8$, 12 and 16 predictor variables. 30 of the sets were such that the predicted variable y was defined as a linear function of $p/2$ predictor variables only and an additional error term e:

$$y = b_0 + b_1x_1 + \ldots + b_{p/2}x_{p/2} + e\,.$$

Other 30 data sets were obtained assuming a specific dependence structure of y from $x_1, \ldots, x_p$: The covariance matrix of the considered variables was constructed from Helmert matrices with an additional condition that the

TABLE 1

Mean CPU times (in minutes) for finding an ε-optimal or optimal regression subset of size k out of p variables recorded on an ODRA 1305 computer.

ε-opt: time needed by the ε-optimal algorithm
opt bb: time needed by the branch and bound algorithm [1]
opt tr: time needed by the traditional all-subset search [2]

p	k	$1 \div 4$	$5 \div 7$	$8 \div 11$	$12 \div 15$
	A. Data sets obtained by the first method (sum of $p/2$ variables)				
$p = 8$	ε-opt	0.019	0.021	–	–
	opt bb	0.018	0.017	–	–
	opt tr	0.030	0.030	–	–
$p = 12$	ε-opt	0.106	0.037	0.055	–
	opt bb	0.110	0.045	0.047	–
	opt tr	0.110	0.330	0.210	–
$p = 16$	ε-opt	0.452	0.748	0.152	0.145
	opt bb	0.521	1.309	0.411	0.134
	opt tr	0.370	4.390	9.130	3.020
	B. Data sets obtained by the second method (Helmert matrices)				
$p = 8$	ε-opt	0.026	0.021	–	–
	opt bb	0.026	0.018	–	–
	opt tr	0.030	0.030	–	–
$p = 12$	ε-opt	0.141	0.220	0.055	–
	opt bb	0.142	0.320	0.052	–
	opt tr	0.110	0.330	0.210	–
$p = 16$	ε-opt	0.615	3.029	1.551	0.145
	opt bb	0.613	3.454	2.943	0.148
	opt tr	0.370	4.390	9.130	3.020

importance of subsequent predictors $1, \ldots, p$ decreases with i $(1 \leq i \leq p)$, the number of the predictor. (For details see [1].)

In our evaluations we assumed $\varepsilon = 0.05$. Moreover, we have used the ratio RSS/SST as the criterion. The algorithm was implemented as a procedure in Algol 1900 and imbedded into a general program contained in the SABA package [2] working on ODRA 1305 computers. The times of run are shown in Table 1. We show there the average time needed in 10 (similar in size) data sets to reach an ε-optimal subset. For comparison we also show in Table 1 the mean times needed in the same data sets to obtain optimal subsets using the branch and bound algorithm described in [1]. We also show the times needed when using a classical algorithm evaluating RSS for all subsets. The times shown in Table 1 are CPU times picked up from standard output obtained on the ODRA 1305 computer when using the GEORGE 3 operating system.

One can see that, generally, for subsets of size $k = 1 \div 4$ the gain, if any, is not big. One might say that in this case the algorithm works very fast and the difference in the performance of the optimal and ε-optimal algorithm is not clearly visible.

When considering subsets of size $k = 5 \div 7$ we should additionally take into account p, the number of predictor variables. For $p = 8$ and $p = 12$ still both algorithms work very fast and'the run times of both algorithms are practically the same (especially, after taking into account the accuracy of the time count). We note in set A for $p = 12$ a very remarkable difference between the two branch and bound algorithms as compared with the traditional all subset search. For $p = 16$ we note a remarkable gain of time, especially for subsets of size $k = 5 \div 7$ and $k = 8 \div 11$.

An example of application of the ε-optimal algorithm in the context of a discriminant analysis carried out in medical data may be found in [3].

References

[1] A. Bartkowiak, *Experience in computing optimal regression by branch and bound*, Zastos. Mat. 20 (1989), 75–86.

[2] —, *SABA, An Algol package for statistical data analysis on the ODRA 1305 computer*, Universitas Wratislaviensis, Wrocław 1984.

[3] A. Bartkowiak, S. Łukasik, K. Chwistecki and M. Mrukowicz, *Search for most discriminative features for IHD considering optimal and ε-optimal subsets yielded by a branch and bound method*, Modelling, Simulation & Control C 12 (1988), 31–39.

ANNA BARTKOWIAK
INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW
UL. PRZESMYCKIEGO 20
51-151 WROCŁAW, POLAND

Received on 3.8.1989;
revised version on 1.10.1991