SEARCH FOR THE MOST DISCRIMINATIVE FEATURES OF CORONARY HEART DISEASE USING A BRANCH AND BOUND METHOD

A. Bartkowiak, S. Łukasik, K. Chwistecki and M. Mrukowicz

Institute of Computer Science, University of Wrocław, and
Cardiological Institute, Medical Academy
Wrocław, Poland

THE MEDICAL PROBLEM

Data have been collected in the Wrocław Coronary Heart Disease (CHD) Prevention Study, being part of the ERICA programme co-ordinated by the WHO Collaborating Centre in Heidelberg. After 4 years of follow-up a sample from the monitored cohort of men working in industrial plants of Wrocław was taken. It comprised 2433 men, of whom 39 showed CHD symptoms. The number of variables considered is p = 12: age, body weight, systolic and diastolic blood pressure, measured casually and in standard conditions, number of cigarettes smoked daily, cholesterol, high density lipoprotein (HDL), triglyceride, uric acid and glucose. The goal is to find variables which differentiate between men with and without CHD symptoms. The task is performed using methods of discriminant analysis. Linear (in p variables) and quadratic (in p + p variables) discriminant functions have been used. A subset of variables is sought which best discriminates between the two groups of men (with and without CHD symptoms). As the criterion of discriminative power the Mahalanobis distance is adopted.

THE BRANCH AND BOUND METHOD IN THE SEARCH FOR THE OPTIMAL SUBSET

Discrimination between two groups can be performed using regression methods (see, for example, Lachenbruch, 1975). Also the Mahalanobis distance can be evaluated as the residual variance in a special case of normal equations (see, for example, Bartkowiak, 1984). Therefore the methods of search for an optimal subset in regression analysis can be transferred directly to the search for an optimal set in discriminant analysis.

In this case, the algorithm used is one elaborated recently by Bartkowiak (1987). It permits the finding of the subset yielding the minimal residual sum of squares in a linear regression problem. The task is to find the subset yielding the maximal Mahalanobis distance. It can be shown (see, for example, Bartkowiak, 1984, p. 59) that the Mahalanobis distance can be obtained as a residual sum of squares after reversing the sign of this residual sum of squares. Therefore the problem of finding the subset with the largest Mahalanobis distance is equivalent to the problem of finding the subset with the smallest residual sum of squares in a properly set linear regression problem. The method of building the regression equation is described, for instance, by Lachenbruch (1975), or Bartkowiak (1984).

Let us define an artificial predictor variable y

$$y = \begin{cases} \dfrac{n_2}{n_1 + n_2} & \text{for individuals belonging to the first group of data comprising } n_1 \text{ individuals} \\ -\dfrac{n_1}{n_1 + n_2} & \text{for individuals belonging to the second group of data comprising } n_2 \text{ individuals} \end{cases} \quad (1)$$

and consider the regression

$$y = b_0 + b_1 x_1 + \ldots + b_p x_p + e \quad (2)$$

To find the optimal subset we proceed as follows:

First, a new order of the variables 1, 2, ..., p is introduced according to the magnitude $Q(-1), \ldots, Q(-p)$ of their residual sums of squares $Q(-i)$, $i = 1, \ldots, p$, defined as the residual sum of squares obtained when introducing all, but the i^{th}, variables into the regression set. Ordering the set of $Q(-i)$, the following sequence is obtained:

$$Q(-1) \geq Q(-2) \geqslant \ldots \geqslant Q(-p) \quad (3)$$

The variables are then relabelled so that their $Q(-i)$s satisfy the inequalities (1).

Next, the search for the best subset is performed considering k + 1 "branches" of generated subsets of size k:

The $(k + 1)^{th}$ branch comprises only one set: 1, 2, ..., k
The k^{th} branch comprises the integers 1, 2, ..., k - 1 and one of the integers p - k + 1, ..., p
The $(k - 1)^{th}$ branch comprises the integers 1, 2, ..., k - 2 and two of the integers p - k + 2, ..., p
... and so on ..., up to
The 1^{st} branch which comprises all k-tuples that can be chosen from the integers 2, 3, ..., p.

It follows from the definition of the subsets comprising a branch that the number of subsets contained in the j^{th} branch is equal to $\binom{p-j}{k-j+1}$:

$$N_j = (\text{number of subsets in the } j^{th} \text{ branch} = \binom{p-j}{k-j+1} \quad (4)$$

Seeking the optimal subset we proceed as follows:

(i) We start from the $(k+1)^{th}$ branch and evaluate the residual sum of squares RSS for this subset. If $RSS < Q(-k)$, we have found the optimum subset and the search is terminated. Otherwise we retain RSS as RSS_o and we proceed for $j = k, k - 1, \ldots, 1$ considering the subsequent "branches".

(ii) For the j^{th} branch we ask whether the current RSS_o is smaller than $Q(-j)$. If yes, then the subset yielding RSS_o is the optimal subset and we stop the search. Otherwise, we generate sequentially the subsets belonging to this branch. Finding an RSS smaller than RSS_o the current RSS is relabelled as RSS_o.

(iii) Having considered all subsets belonging to the j^{th} branch, we diminish j by one. If $j > 0$, we start (ii), otherwise RSS_o is the minimal value, and the subset which yielded RSS_o is the optimal subset.

THE CONCEPT OF AN ε-OPTIMAL SUBSET

The residual sum of squares $Q(-i)$ defined by (1) is a lower bound for the RSSs evaluated for all subsets belonging to the i^{th} branch and simultaneously for branches with a smaller index than i. If for a particular subset found in the $(i+1)^{th}$ branch the RSS is smaller than $Q(-i)$, then no subset belonging to the branches 1, 2, ..., i can give a smaller RSS.

Instead of considering the inequality

$$RSS < Q(-i) \tag{5}$$

we could verify another inequality

$$RSS < Q(-i) + \varepsilon \tag{6}$$

In practice we consider instead of (6) a modified inequality

$$RSS < Q(-i) + \varepsilon SS(y), \tag{6a}$$

where $SS(y)$ is the total adjusted sum of squares of the variable y introduced in (1):

$$SS(y) = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad n = n_1 + n_2 \tag{7}$$

A subset with RSS satisfying (6) or (6a) is said to be ε-optimal (its RSS differs from the RSS of the optimal subset not more than by ε).

RESULTS FOR THE CHD DATA

We consider first p = 12 and next p = 24 variables with the artificial regression functions (see (2))

$$y = b_0 + b_1x_1 + \ldots + b_{12}x_{12} + e, \tag{8a}$$

$$y = b_0 + b_1x_1 + \ldots + b_{12}x_{12} + b_{13}x_1^2 + \ldots + b_{24}x_{12}^2 + e \tag{8b}$$

The two regressions lead appropriately to linear and quadratic discriminant functions, respectively. We have been looking for optimal subsets of size k = 3, 4, 5. The total number of subsets to be evaluated when using the traditional all-subset search is given in Table 1. The branches and the N_js, the number of subsets in these branches, are given in Table 2 (the N_js were evaluated using (4)).

Table 1. The Number of Subsets to be Evaluated in the All-subset Search. (p - number of variables, k - size of the subset)

k	p = 12	p = 24
3	$\binom{12}{3} = 220$	$\binom{24}{3} = 2024$
4	$\binom{12}{4} = 495$	$\binom{24}{4} = 10626$
5	$\binom{12}{5} = 792$	$\binom{24}{5} = 42504$

It can be seen that the totals of the numbers of subsets in the branches numbers $k+1$, k, $k-1$, ..., 1 are equal to the numbers of subsets given in Table 2.

Table 2. Branches and the Numbers of Subsets in These Branches When Seeking Subsets of Size $k = 3, 4, 5$ Out of $p = 12$ and $p = 24$ Variables

Number of the branch j	Considering $p = 12$ Variables N_j - number of subsets	Considering $p = 24$ Variables N_j - number of subsets
	size of the subset: $k = 3$	
$j = 4$	$\binom{8}{0} = 1$	$\binom{20}{0} = 1$
$j = 3$	$\binom{9}{1} = 9$	$\binom{21}{1} = 21$
$j = 2$	$\binom{10}{2} = 45$	$\binom{22}{2} = 231$
$j = 1$	$\binom{11}{3} = 165$	$\binom{23}{3} = 1771$
Total	220	2024
	size of the subset: $k = 4$	
$j = 5$	$\binom{7}{0} = 1$	$\binom{19}{0} = 1$
$j = 4$	$\binom{8}{1} = 8$	$\binom{20}{1} = 20$
$j = 3$	$\binom{9}{2} = 36$	$\binom{21}{2} = 210$
$j = 2$	$\binom{10}{3} = 120$	$\binom{22}{3} = 1540$
$j = 1$	$\binom{11}{4} = 330$	$\binom{23}{4} = 8855$
Total	495	10626
	size of the subset: $k = 5$	
$j = 6$	$\binom{6}{0} = 1$	$\binom{18}{0} = 1$
$j = 5$	$\binom{7}{1} = 7$	$\binom{19}{1} = 19$
$j = 4$	$\binom{8}{2} = 28$	$\binom{20}{2} = 190$
$j = 3$	$\binom{9}{3} = 84$	$\binom{21}{3} = 1330$
$j = 2$	$\binom{10}{4} = 210$	$\binom{22}{4} = 7315$
$j = 1$	$\binom{11}{5} = 462$	$\binom{23}{5} = 33649$
Total	792	42504

Table 3. Results of the Search for an ε-Optimal and Optimal Subset Considering p = 12 Variables

k, Size of the Subset	Branches Considered	Subset Chosen (notation as above)	Fraction of Subsets considered
ε-optimal subset with ε = 0.02 according to (6a)			
3	j = 3	1, 11, 12	10:220 = 0.045454
4	j = 4	1, 8, 11, 12	9:495 = 0.018182
5	j = 5	1, 4, 6, 11, 12	8:792 = 0.010010
optimal subset			
3	j = 3	1, 11, 12	10:220 = 0.045454
4	j = 4 v j = 3	1, 8, 11, 12	45:495 = 0.090909
5	j = 5 v j = 4	1, 4, 6, 11, 12	36:792 = 0.045454

From Table 3 it can be seen that the ε-optimal sets and optimal sets comprise the same variables. Generally, the optimal algorithm worked for p = 12 variables very rapidly and the difference in time between the two algorithms is not very large.

When considering p = 24 variables the ε-optimal subset was found very speedily. When seeking the optimal subset we were not lucky and had to evaluate all subsets. The subset of size k = 3 is the same for both methods. The subsets of size k = 4 and k = 5 differ by one variable. Performing the calculations for p = 24 variables and seeking a good subset, the great advantage of searching for a ε-optimal subset can be seen: the gain in time here was tremendous! The details are given in Table 4.

Table 4. Results of the Search for an ε-Optimal and Optimal Subset Considering p = 24 Variables

k, Size of the Subset	Branches Considered	Subset Chosen (notation as above and (8b))	Fraction of Subsets considered
ε-optimal subset with ε = 0.02 according to (6a)			
3	j = 3	1, 12, 23	22: 2024 = 0.010870
4	j = 4	1, 4, 12, 23	21:10626 = 0.001976
5	j = 5	1, 4, 12, 23, 24	20:42504 = 0.000471
optimal subset			
3	j = 3 v j = 2 v j = 1	1, 12, 23	all = 1.0
4	j = 4 v j = 3 v j = 2 v j = 1	1, 12, 23, 24	all = 1.0
5	j = 5 v j = 4 v j = 3 v j = 2 v j = 1	1, 11, 12, 23, 24	all = 1.0

DISCUSSION OF THE RESULTS

The method described shows a great advantage over the traditional one. The application of the branch and bound method for discriminant analysis was possible only because the Mahalanobis distance, taken as the criterion of the discriminative power of variables under investigation, has the property of being a monotonic function of the number of variables in the discriminative set being considered: adding a new variable to this set, the Mahalanobis distance between the two groups of data can never be decreased, but it may possibly increase.

The results obtained for the detailed medical problem are, at first sight, a little surprising. The most discriminative features here are: age, uric acid and glucose. These (except perhaps for age) are not judged as the most important risk factors for Coronary Heart Disease (Multiple Risk Factor Intervention Trial Research Group, 1982). In particular, the variable "uric acid" is somehow questionable, although there are some reports on the importance of this variable when considering CHD (Persky et al., 1979). It is surprising that the feature "smoking" was not revealed by the search procedure. One possible explanation could be that the question, "How many cigarettes do you smoke per day?", was not precise enough and did not give much information on the smoking history of the interviewed individual. Another possibility is that the groups of data considered are truly not differentiated at all and that the results obtained are spurious.

It should be remembered that the people considered were still in the working age group and were employed in industrial plants. It can be concluded that the CHD symptoms stated in these people (men) were not very advanced because they were still able to carry out their professional activities. Therefore, the final (medical) conclusion that these people with stated CHD symptoms do not differ statistically from those with no CHD symptoms - at least with regard to the 12 parameters considered - is not surprising.

REFERENCES

Bartkowiak, A., 1984, "SABA - An Algol Package for Statistical Data Analysis on the ODRA 1305 Computer", Universitas Wratislaviensis, Wrocław.

Bartkowiak, A., 1987, Experience in computing optimal regression by branch and bound, Zastosowania Matematyki/Applicationes Mathematicae, 20(2): (in press).

Lachenbruch, P., 1975, "Discriminant Analysis", Hafner Press, Macmillan, London.

Multiple Risk Factor Intervention Trial Research Group, 1982, Multiple risk factor intervention trial. Risk factor changes and mortality results, J.A.M.A., 248:1485.

Persky, V. W., Dyer, A. R., Idris-Soren, E., Stamler, J., Shekelle, R. B., Schoenberger, J. A., Berksom, D. M., and Lindberg, H. A., 1979, Uric acid: a risk factor for coronary heart disease?, Circulation, 59:969.

Advances in Biomedical Measurement

Edited by

Ewart R. Carson
City University
London, England

Peter Kneppo
Slovak Academy of Sciences
Bratislava, Czechoslovakia

and

Ivan Krekule
Czechoslovak Academy of Sciences
Prague, Czechoslovakia

PLENUM PRESS • NEW YORK AND LONDON

Library of Congress Cataloging in Publication Data

IMEKO Conference on Advances in Biomedical Measurement (4th: 1987: Bratislava, Czechoslovakia)
Advances in biomedical measurement.

"Proceedings of the Fourth IMEKO Conference on Advances in Biomedical Measurement, held May 1987, in Bratislava, Czechoslovakia"—T.p. verso.
Includes bibliographies and index.
1. Medical electronics—Congresses. 2. Human physiology—Measurement—Congresses. 3. Imaging systems in medicine—Congresses. 4. Expert systems (Computer science)—Congresses. 5. Biomedical engineering—Instruments—Congresses. I. Carson, Ewart R. II. Kneppo, Peter. III. Krekule, Ivan. IV. International Measurement Confederation. V. Title. [DNLM: 1. Biomedical Engineering—congresses. 2. Electronics, Medical—congresses. W3 IM309 4th 1987a / QT 34 I32 1987a]
R895.A2I43 1987 610'.28 88-12533
ISBN 0-306-42923-3

Proceedings of the Fourth IMEKO Conference on Advances in Biomedical Measurement, held May 1987, in Bratislava, Czechoslovakia

Printed in the United States of America