An Application of Linear Programming in Cluster Analysis

PIOTR MARIAN WISNIEWSKI , IRMA LÓPEZ SAURA , GABRIEL VELASCO SOTOMAYOR
Departamento de Matemáticas
TEC de Monterrey, Campus Ciudad de México
Calle del Puente No.222, C.P.14380 México, D.F.
MEXICO

Abstract: A method is developed with a view to finding a solution to a certain type of cluster analysis problems by means of linear programming. In order to achieve the optimal solution, a partition of a set of objects into disjoint subsets is used. To this effect the concept of "natural number of groups" is introduced, whereby an improvement is accomplished as regards both the methodology and the solution. Previous relevant efforts in this connection are discussed too.

Key-Words – Cluster analysis , overlapping groups , linear programming, partition, "natural" number of groups

1 Introduction

The partition of a set of multidimensional objects into nonempty homogeneous subsets (groups) is the central problem of cluster analysis. The application of various measures of homogeneity to create cluster leads, more or less, to different refined techniques of clustering. The use of computers was conducive to the development of methods which allow us to select the best partitions from all, or almost all, possible partitions. To this category belong those methods based on the use of mathematical programming for cluster seeking.

Starting from the optimal partition of a set of N multidimensional objects into M disjoint subsets obtained by linear programming with a minimum sum of squares within groups as the objective function, a procedure for seeking overlapping groups that utilizes values of dual variables has been developed. Identification of the subset is obtained by solving a linear program with parameterized right hand sides. The parameterization is controlled by information extracted from the solution of the dual program. In this sequential procedure in each step a minimal increase of the value of the objective function is assured. A "natural" number of groups has also been defined.

2 Problem Formulation

Suppose that a set of N multidimensional objects represented by a $p \times N$ matrix of observations $\underline{Y} = (y_{ij})$, $i = 1,2,...,p$; $j = 1,2,...,N$ is given. In a Euclidean classification such a set of N objects can be represented by N points in the p-dimensional Euclidean space. The sum of squares criterion (*SSC*) defines the optimal partition of the points into M disjoint groups as that which minimizes the total within-group sum of squared distance from the M centroids.

Rao [6] proved that the solution of the problem of the partition of the set N of objects into M disjoint groups, by *SSC* , can be obtained as a solution to the following linear program:

(1) $$\min_{\underline{x}} \underline{c}' \underline{x} \quad ,$$

with the constrains

$$\begin{pmatrix} \underline{A} \\ \underline{1}' \end{pmatrix} \underline{x} = \begin{pmatrix} \underline{1} \\ M \end{pmatrix},$$

where $\underline{x}$ is a $\left(2^N - 1\right) \times 1$ binary vector and where the k-th component of the vector $\underline{c}$ is

$$c_k = \sum_{j \in I(G_k)} \sum_{i=1}^{p} (y_{ij} - \bar{y}_{ik})^2 \quad ,$$

$I(G_k)$ being a set of indexes of objects belonging to the subset G_k, with $\bigcup_k I(G_k) = 1,2,...,N$ for $k = 1,2,...,2^N - 1$, and $\bar{y}_{ik}$ denoting the i-th coordinate of the centroid of the k-th group, and, moreover, where $\underline{A}$ is an $N \times \left(2^N - 1\right)$ binary matrix constructed in such a way that $a_{jk} = 0$ if the j-th

object does not belong to this group and $a_{jk} = 1$ if the object belongs to it, $\underline{1}$ is a vector of an appropriate size which consists of ones only, and $\underline{x}$ is a $\left(2^N - 1\right) \times 1$ vector of unknown quantities.

Using the result of Gordon and Henderson [2], Harabasz and Wisniewski [3] proved that the program (1) always has a binary solution. This fact allows to find a solution of the above mentioned program by means of standard procedures for linear programs with nonnegative real variables. The procedure of the formulation of the program (1), utilizing the generalized string property (Rao[6]) to reduce the dimension of the matrix $\underline{A}$, was also discussed by Harabasz and Wisniewski [3]. Let us note that the problem of the partition of a set into disjoint subsets, under an appropriately selected criterion of optimality, is known in mathematical programming as the problem of set partition.

Let us formulate a non-symmetrical dual problem for program (1) :

(2) $$\max_{\underline{u}} \left(\underline{1}^{/} \,\vdots\, M\right)\underline{u}$$

with the constrains

$$\left(\underline{A}^{/} \,\vdots\, \underline{1}\right)\underline{u} \le \underline{c}$$

where $\underline{u}$ is an $(N+1)$ vector of real variables. The solution of the program (2) can be used to find, at the same time, the "natural" number of groups and the overlapping groups.

Also, it is worth mentioning that the solution of the program (1) and that of the program

$$\min_{\underline{x} \ge \underline{0}} \underline{c}^{/} \underline{x}$$

with the constraints

$$\begin{cases} \underline{A}\underline{x} \ge \underline{1} \\ \underline{1}^{/} \underline{x} = M \end{cases}$$

are identical.

Thus, the solution of the dual program (2) and that of the program

$$\max_{\underline{u}} \left(\underline{1}^{/} \,\vdots\, M\right)\underline{u}$$

with the constrains

$$\left(\underline{A}^{/} \,\vdots\, \underline{1}\right)\underline{u} \le \underline{c}$$

where $u_1, u_2, \ldots, u_N$ and u_{N+1} is unrestricted in sign are identical.

3 Problem Solution

3.1 The determination of the "natural" number of groups

The number of cluster, M, is usually an unknown quantity. The majority of known cluster algorithms require the specification of the number M as known. This requirement is in many cases unnatural. We will show that the value of the dual variable u_{N+1} in program (2) can be used to determine the natural number of groups. Let us establish the sequence of program (1) for $M = 2, 3, \ldots N-1$ and let us denote by $\{f(M)\}$, $M = 2, 3, \ldots N-1$, the sequence of the minimal values of the objective function for those programs. Furthermore, let $\{u_{N+1}(M)\}$ be the sequence of values of the dual variables fulfilling the following requirement $\underline{1}^{/} \underline{x} = M$. Of importance is now the following

Theorem : The sequence $\{u_{N+1}(M)\}$ for $M = 2, 3, \ldots N-1$, is nondecreasing.

Proof : Let $\underline{B}_M$ be the basis of the optimal solution of the program (1) for the division into M groups. With the use of this basis, values of the solution of programs (1) and (2) will be as follows :

$$\underline{x}_M = \underline{B}_M^{-1} \begin{pmatrix} \underline{1} \\ \hline M \end{pmatrix} \quad , \quad \underline{u}_M = \left(\underline{B}_M^{-1}\right)^{/} \underline{c}$$

where $f(M) = \underline{c}^{/} \underline{x}_M = \left(\underline{1}^{/} \,\vdots\, M\right)\underline{u}_M$. If M is increased by 1, two tendencies may occur :

1) The basis $\underline{B}_M$ is still the basis of the optimal solution of (1) for $M+1$. Then

$$f_{\underline{B}_M}(M+1) = \underline{c}^{/} \underline{x}_{M+1} = \underline{c}^{/} \underline{B}_M^{-1} \begin{pmatrix} \underline{1} \\ \hline M+1 \end{pmatrix} =$$

$$= \underline{c}^{/} \underline{B}_M^{-1} \begin{pmatrix} \underline{1} \\ \hline M \end{pmatrix} + \underline{c}^{/} \underline{B}_M^{-1} \begin{pmatrix} \underline{0} \\ \hline 1 \end{pmatrix} = f_{\underline{B}_M} + u_{N+1}(M) \quad ,$$

i.e. $f_{\underline{B}_M}(M+1) - f_{\underline{B}_M}(M) = u_{N+1}(M)$.

2) The basis $\underline{B}_M$ is not the basis of the optimal solution of (1) for $M+1$. Let the basis of the optimal solution be $\underline{B}_{M+1} \ne \underline{B}_M$. Then

$$f_{\underline{B}_{M+1}}(M+1) = \underline{c}^{/}\,\underline{B}_{M+1}^{-1}\begin{pmatrix}\underline{1}\\ \hline M+1\end{pmatrix} =$$

$$= \underline{c}^{/}\,\underline{B}_{M+1}^{-1}\begin{pmatrix}\underline{1}\\ \hline M\end{pmatrix} + \underline{c}^{/}\,\underline{B}_{M+1}^{-1}\begin{pmatrix}\underline{0}\\ \hline 1\end{pmatrix},$$

$$f_{\underline{B}_{M+1}}(M+1) \le f_{\underline{B}_M} + u_{N+1}(M)\ ,$$

i.e.

$$f_{\underline{B}_M+1}(M+1) - f_{\underline{B}_M}(M) = u_{N+1}(M+1)\ .$$

Thus

$$u_{n+1}(M+1) \ge f(M+1) - f(M) \ge u_{N+1}(M)\ ,$$

with the equality on the left when $f(M) = f_{\underline{B}_{M+1}}(M)$ and with that on the right when $f(M+1) = f_{\underline{B}_M}(M+1)$.

Hence $u_{N+1}(M+1) \ge u_{N+1}(M)$ and the proof is complete.

Based on the above characteristic of the set $\{u_{N+1}(M)\}$ we shall introduce the definition of the "natural" partition of objects into M_0 groups.

Definition : The partition of a set of *N* objects into M_0 groups is called natural if M_0 is the smallest positive integer for which

$$\min_{M\in\{2,3,\ldots,N-1\}}\{u_{N+1}(M+1) - u_{N+1}(M)\} =$$

$$= u_{N+1}(M_0+1) - u_{N+1}(M_0)\ .$$

It should be pointed out that, for practical purposes, a sequence of solutions of the program (1), where *M* changes from 2 to $N-1$, can be achieved throughout the parameterization of the $(N+1)$-th component of the vector on the right hand side of the constraints.

3.2 Overlapping Groups

The optimal solution of the program (1) indicates the partition of a set of objects into disjoint subsets. In many classification problems the requirement that the subsets obtained are disjoint is unnatural and may be contradictory to possible significant relations between the objects. In such a situation procedures for finding overlapping subsets are desirable. The majority of known procedures are based on the application of conventional threshold values for either the measure of similarity or of dissimilarity (distance) between objects (Bock[1]). It is also possible to control the number of objects belonging to the common overlapping part, i.e. to the cross section of clusters (Jardine and Sibson [4]).

Starting from the linear program (1) we can see that subsets in its solution are disjoint because every object may occur in a subset only once. This is ensured by requirements $\underline{A}\,\underline{x} = \underline{1}$. The change of the right hand side in the *j*-th requirement ($j = 1, 2, \ldots, N$) of this set of equations from 1 to another natural number $1+\beta_j$ causes the occurrence of the *j*-th object in $1+\beta_j$ subsets. It is obvious that $1+\beta_j$ should not be greater than the number of groups *M*. The occurrence of the *j*-th object in more than one group will cause a significant increase in the minimum values of the objective function. The objects ($j^{/}s$) and the ($\beta_j^{/}s$) should be chosen in such a manner that the increase in the value of the objective function in the minimum solution is the lowest. The lowest increase in the value of the objective function occurs for $\beta_j = 1$.

The index *j*, with the same criterion, can be chosen with the use of values of dual variables of the program (2). Number *j* is the index of the dual variable of the lowest value. The smallest possible increase in the value of the objective function is u_j if the new minimum solution is achieved for the same basis.

Starting from the values of the dual solutions of the program (2) we choose the row index with a set of constraints where the right hand side will be increased by one. Then we obtain a sequence, $v = 1, 2\ldots$, of linear programs

(3) $$\min_{\underline{x}} \underline{c}^{/}\,\underline{x}$$

with a set of constraints

$$\begin{cases}\underline{A}\,\underline{x} = 1 + \beta(v)\\ \underline{1}^{/}\,\underline{x} = M\end{cases}$$

where at every stage a certain coordinate of the vector $\underline{\beta}$ increase by one without exceeding $M-1$, e.g. $\underline{\beta}(v) = \underline{\beta}(v-1) + \underline{e}_{j_{v-1}}$, where $\underline{e}_{j_{v-1}}$ is the vector with one occupying the position j_{v-1}.

The relevant sequence of dual programs is

(4) $$\max_{\underline{u}}\left(\underline{1}^{/} + \underline{\beta}(v) \;\vdots\; M\right)\underline{u}\ ,$$

with the constraints

$$\left(\underline{A}^{/} \;\vdots\; \underline{1}\right)\underline{u} \le \underline{c}\ .$$

The formulation of the following programs (3) is obtained by parameterization of right hand sides of the program (1), where the choice of vectors $\underline{e}_j$ is determined by the dual solutions. In order to generate the following programs (3) it is necessary to specify the rule terminating this process, e.g. the maximum value of v . The following rule seems advisable : *"The process of the formulation of overlapping groups should be terminated in such a manner that the total within-group sum of squares does not exceed the total sum of squares"*.

4 Conclusion

Let us analyze an example given by Rao [5] concerning the partition of $N = 12$ Indian castes in relation to $p = 9$ anthropometric traits. The castes are: Brahmin {Basti B_1}, Brahmin {Other B_2}, Chattri {Ch}, Muslim {M}, Bhatu $\{C_1\}$, Habru $\{C_2\}$, Bhil {Bh}, Dom {D}, Ahir $\{A_1\}$, Kurmi $\{A_2\}$, Other Artisan $\{A_3\}$, Kahar $\{A_4\}$. Our purpose is to find a natural number of groups on the basis of rules given in 3.1. This is why, for $M = 2, 3, \ldots, 11$ the linear programs (1) and the corresponding dual programs (2) have been generated. The results are as follows:

Table 1
Values of the dual variable $u_{13}(M)$, objective function and the increase in objective function for $M = 2, 3, \ldots, 11$

M	$u_{13}(M)$	$f(M)$	$f(M+1) - f(M)$
2	-3.2877	10.0389	-3.2877
3	-2.4891	6.7521	-2.4891
4	-2.2677	4.2630	-1.8330
5	-0.9725	2.4300	-0.6600
6	-0.6050	1.7700	-0.5750
7	-0.5567	1.9550	-0.5567
8	-0.5567	0.6383	-0.2433
9	-0.2433	0.3950	-0.2000
10	-01350	0.1950	-0.1350
11	-0.1350	0.0600	-0.0600

It can be seen that

$$\min_{M \in \{2,3,\ldots,11\}} \{u_{N+1}(M+1) - u_{N+1}(M)\} =$$
$$= u_{13}(8) - u_{13}(7) = 0 .$$

Thus, according to our definition, the partition of $N = 12$ castes into $M_0 = 7$ groups is a natural partition. When $M = 7$ changes into $M = 8$ there is a definite increase in the sequence $\{u_{13}(M)\}$. Thus, $M = 7$ as a "natural" number of groups seems to be justified because the formation of $M = 8$ groups causes the minimum increase in the value of objective function. Solution of the programs (1) and of corresponding programs (4) led to the following overlapping clusters of 12 Indian castes :

$\{B_1, B_2, A_1, A_2, A_3\}$, {M , Ch} , {Bh , D} , $\{A_1, A_2, A_3, A_4\}$,$\{C_1, C_2\}$, $\{A_2, A_3, A_4, M\}$, $\{C_2, A_1\}$.

References :

[1] H.H. Bock, Automatische Klassifikation, Vandenhoeck and Ruprecht, Göttingen, 1974

[2] A.D. Gordon and J.J. Henderson, An algorithm for Euclidean sum of squares classification, Biometrics No.33, 1977, pp.355-361

[3] J.S. Harabasz and Piotr Wisniewski, Generation of a linear program for grouping problem, Roczn. AR w Poznaniu, Algor. Biom i Stat, No 11, 1984, p.137.145

[4] N. Jardine and R. Sibson, Mathematical Taxonomy, Wiley, New York, 1971

[5] C.R. Rao, Advavced Statistical Methods in Biometrics, Wiley, New York, 1952

[6] M.R. Rao, Cluster analysis and mathematical programming, J.Amwr. Statist. No 66, 1971, pp. 622-626