Control
and Cybernetics
VOL. 15 (1986) No. 2

Anticlustering: Maximizing the variance criterion

by

H. SPÄTH

Fachbereich Mathematik
Universität Oldenburg
Postfach 2503
2900 Oldenburg, F. R. Germany

In certain applications a set of m objects characterized by values of s variables x_{ij} ($i=1, ..., m$; $j=1, ..., s$) is to be split into n subsets that are as similar as possible. It is shown that the variance criterion, i.e., the search for a partition $C_1, ..., C_n$ of $\{1, ..., m\}$ for which

$$\sum_{j=1}^{n} \sum_{i \in C_j} \|x_i - \bar{x}_j\|^2$$

attains its maximum value, provides a suitable formulation. A heuristic solution method and numerical results for several examples are given.

1. Problem

Usually the aim of cluster analysis is to split a set of m given objects into groups (clusters) of similar objects that are different from each other as possible. Within a practical application for filling stations as objects (see Braun (1978) for a related problem) another objective appeared. The objects were to be split into groups (anticlusters) of dissimilar objects that could be as similar as possible.

Assuming that the objects are characterized by values

$$((x_{ik},\ i=1, ..., m),\ k=1, ..., s) \tag{1}$$

for s quantitative variables, the most popular objective function for the first purpose is the minimum variance criterion. A partition $C=(C_1, ..., C_n)$ of length $1<n<m$ of $\{1, ..., m\}$ is sought such that the average value of variances over all clusters C_j attains a minimum:

$$\min_{C} \sum_{j=1}^{n} \sum_{i \in C_j} \|x_i - \bar{x}_j\|^2, \tag{2}$$

where

$$x_i=(x_{i1}, \ldots, x_{is})^T,$$

$$\bar{x}_j=\frac{1}{|C_j|}\sum_{i\in C_j} x_i \tag{3}$$

and $\|\cdot\|$ denotes the Euclidean norm.

Due to the well-known relation, Späth (1985):

$$\sum_{i=1}^{m}\|x_i-\bar{x}\|^2=\sum_{j=1}^{n}\sum_{i\in C_j}\|x_i-\bar{x}_j\|^2+\sum_{j=1}^{n}|C_j|\,\|\bar{x}_j-\bar{x}\|^2, \tag{4}$$

where

$$\bar{x}=\frac{1}{m}\sum_{i=1}^{m} x_i,$$

and as the left hand side of (4) does not depend on the partition C, it is clear that (2) is equivalent to

$$\max_C \sum_{j=1}^{n}|C_j|\,\|\bar{x}_j-\bar{x}\|^2. \tag{5}$$

This means that the group means $\bar{x}_j$ are in some sense as far as possible away from the overall mean $\bar{x}$.

Concerning the above mentioned other objective, we are looking for groups as similar as possible. If we accept as reasonable aim that now the average value of variances attains its maximum, i.e.

$$\max_C \sum_{j=1}^{n}\sum_{i=C_j}\|x_j-\bar{x}_j\|^2, \tag{6}$$

then, using (4) again, we have as equivalent goal function

$$\min_C \sum_{j=1}^{n}|C_j|\,\|\bar{x}_j-\bar{x}\|^2. \tag{7}$$

Thus, in an optimal partition the group means $\bar{x}_j$ are now as close as possible in the above sense to the overall mean $\bar{x}$ and, thus, to each other. If the number m of objects is large enough in relation to the number n of groups, then the value of (7) tends to zero, i.e. $\bar{x}_j\to\bar{x}$ for each $j=1, \ldots, n$, and, furthermore, (6) tends to the left hand side of (4), i.e. to the total variance.

2. Solution method

As we will see from the numerical examples given in Section 3, random partitions usually do not approximate (6) very well. Thus, we improve them by suitably modifying the well-known exchange method to obtain an approximate solution the combinatorial optimization problem (2). That method, applied for (2), improves

some (random) initial partition by successively moving on trial each object from its cluster to all the other ones, and by shifting it, if there is any reduction at all, to that one where the first term on the right side of (4) decreases most, otherwise taking the next object and finally passing through all the objects until no further improvement occurs. The whole procedure is repeated for several, say 20, initial partitions and this partition is selected as approximate solution for which the objective function is the lowest. A FORTRAN subroutine TRWEXM and a main program together with numerical examples and a performance test are given in Späth (1985) p. 149 and p. 151–154.

The modification for (6) consists, in changing only three statements of TRWEXM, Späth (1985) as listed below.

Line 24: Replace EQ=BIG by EQ=O.
Line 38: Replace IF (EJ.GE.EQ) ... by IF (EJ.LE.EQ) ...
Line 43: Replace IF (EQ.GE.EP*R) ... by IF (EQ.LE.EP) ...

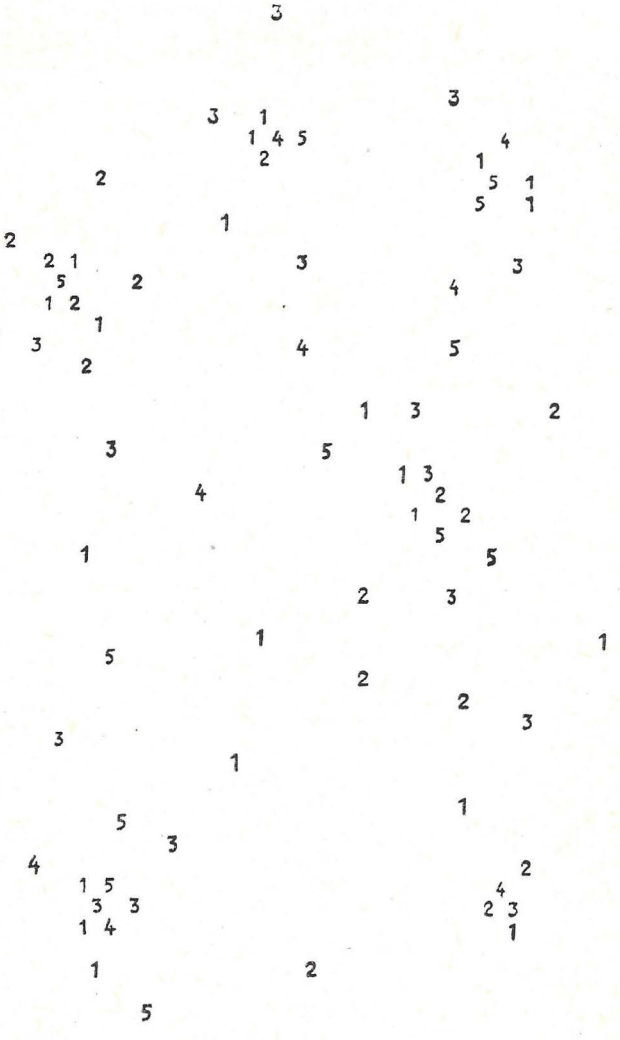
3. Numerical examples

As examples we have taken for simplicity the four data sets with $m=37, 41, 44, 73$ and $s=2$ from Späth (1985), p. 144 which are visualized on p. 146 there. For (2) the results for 20 different random initial partitions are given in Späth (1985), p. 151–54. The results for the same initial partitions but for the objective (6) are

Table 1

Example	n	2	3	4	5	6	7	8	9
1	A	957	886	928	865	843	819	796	765
	B	1068	1038	1016	981	930	905	883	856
	C	1116	1113	1103	1066	1042	1045	1044	1023
	D	1116	1115	1114	1113	1112	1110	1106	1106
	E	1117	1116	1116	1116	1115	1114	1114	1114
2	A	1550	1485	1366	1270	1268	1216	1237	1183
	B	1673	1620	1582	1524	1472	1458	1404	1375
	C	1713	1690	1690	1659	1616	1645	1601	1549
	D	1714	1713	1713	1709	1710	1709	1707	1705
	E	1715	1714	1714	1714	1714	1713	1712	1712
3	A	1189	1187	1090	942	996	900	982	947
	B	1321	1292	1250	1209	1182	1144	1122	1088
	C	1351	1336	1325	1320	1266	1241	1221	1202
	D	1352	1352	1351	1351	1347	1349	1343	1344
	E	1352	1352	1352	1352	1352	1352	1350	1350
4	A	2649	2591	2571	2457	2302	2374	2272	2241
	B	2746	2701	2681	2634	2586	2537	2513	2467
	C	2787	2782	2754	2746	2723	2688	2653	2648
	D	2787	2787	2787	2786	2785	2784	2784	2782
	E	2787	2787	2787	2787	2787	2787	2786	2786

summarized in Table 1. For each example this table contains the minimal (A), average (B), and maximal (C) values of the objective function divided by ten and rounded to four digits for those 20 initial partitions and the minimal (D) and maximal (E) value for the corresponding final partitions after applying the modified exchange method. Table 2 contains, again for the same four examples, the minimal (F), average (G) and maximal (H) percentage gain obtained in this way and calculated



M= 73 N=5 NR= 1 IT= 2 D= 0.2787E+05 (TRW)

1 : (21)	0.8249E+04	2 : (16)	0.5868E+04
3 : (16)	0.6272E+04	4 : (8)	0.3236E+04
5 : (12)	0.4240E+04		

Fig. 1

Table 2

Example	n	2	3	4	5	6	7	8	9
1	F	0.0	0.0	1.1	4.7	7.0	6.6	6.7	8.9
	G	4.6	7.5	9.8	13.8	19.9	23.1	26.2	30.1
	H	16.7	30.0	20.3	29.0	32.2	36.0	39.9	45.6
2	F	0.1	1.4	1.4	3.3	6.1	4.1	6.9	10.5
	G	2.5	5.8	8.3	12.5	16.4	17.5	21.9	24.5
	H	10.6	15.4	25.5	35.0	35.2	40.9	38.4	44.7
3	F	0.0	1.2	2.0	2.4	6.8	8.9	10.6	12.3
	G	2.3	4.6	8.2	11.8	14.4	18.2	20.3	24.1
	H	13.7	13.9	24.0	43.5	35.7	50.2	37.7	42.6
4	F	0.0	0.0	1.1	1.5	2.4	3.7	5.0	5.2
	G	1.5	3.2	4.0	5.8	7.8	9.9	10.9	12.9
	H	5.2	7.6	8.4	13.4	21.1	17.4	22.6	24.3

according to the figures of Table 1. For illustration, Fig. 1 contains the group memberships, for example 4 and $n=5$. (The position of the points (objects) is given by the values of the two variables).

4. Conclusion

As it can be seen from the Tables 1 and 2 the modified exchange method works very well. The values of the final partitions are nearly equal in all cases. (This is different when applying the exchange method for (2)). For $n=2, \ldots, 9$ the values of the objective function decrease very slowly, i.e., the value of (7) is indeed nearly zero all the time. For larger n, of course, this would not have to be that way. Finally the gain in the objective function value as against random partitions is remarkable and is increasing with the number of groups.

References

[1] BRAUN H. Strukturanalyse eines Tankstellennetzes. In: Späth H. Ed. Fallstudien Operations Research. Bd. 1. Munich, R. Oldenburg, 1978.

[2] SPÄTH H. Cluster dissection and analysis. Theory, examples, and FORTRAN programs Chichester, Horwood, 1985.

Antyklustering: kryterium maksymalizacji wariancji

W pewnych zastosowaniach zbiór m obiektów scharakteryzowanych przez wartości s zmiennych x_{ij} $(i=1, \ldots, m, j=1, \ldots, s)$ powinien być rozbity na n możliwie podobnych podzbiorów. Po-

kazano, że w tym przypadku odpowiednim podejściem jest maksymalizacja wariancji, tzn. szukanie takiego rozbicia $C_1, \ldots, C_n$ zbioru $\{1, \ldots, m\}$, dla którego

$$\sum_{j=1}^{n} \sum_{i \in C_j} \|x_i - \bar{x}_j\|^2$$

osiąga wartość maksymalną. Podano heurystyczną metodę rozwiązania tego zagadnienia i przytoczono wyniki obliczeń dla kilku przykładów.

Антикластеризация: критерий максимизации дисперсии

В некоторых приложениях совокупность m объектов характеризируемых значения s переменных x_{ij} $(i=1, \ldots, m, j=1, \ldots, s)$ должна быть разбита на n, возможно сходных подсовокупностей. Показано, что в этом случае надлежащим подходом является максимизация дисперсии, то есть поиск такого разбиеиия $C_1, \ldots, C_n$ совокупности $\{1, \ldots, m\}$, для которого

$$\sum_{j=1}^{n} \sum_{i \in C_j} \|x_i - \bar{x}_j\|$$

достигает максимального значения. Предложено эвристический метод решения этой задачи и приведено результаты вычислений для нескольких примеров.