Indian J. Genet., **56** (4): 412–424 (1996)

A COMPUTER ORIENTED ITERATIVE ALGORITHM FOR CLUSTERING

K. M. SURESH AND V. K. G. UNNITHAN

Department of Agricultural Statistics, College of Horticulture
Kerala Agricultural University, Trichur

(Received: November 11, 1987; accepted: April 2, 1996)

ABSTRACT

A new computer oriented iterative algorithm for formation of clusters using Mahalanobis D^2 values is proposed. The procedure is free from the drawbacks of Tocher's method of clustering using D^2 values, viz., i) the stopping rule for formation of any cluster is arbitrary, and ii) often a genotype belonging to a cluster has on the average a smaller D^2 value with the genotypes of a different cluster than the one it belongs to.

Key words: Mahalanobis D^2, clustering, iterative algorithm.

Mahalanobis D^2 statistic [1] a measure of distance between two populations, taking variation within population also into consideration, is widely used for clustering the genotypes. The procedure now being followed using D^2 was suggested by Tocher [2]. It starts with those two genotypes having minimum value of D^2 and identifies a third genotype which has the smallest average D^2 from the first two. The fourth genotype is chosen which has the smallest average D^2 from the first three and so on. If at any stage the increase in average D^2 for a genotype appears to be higher as compared to the previous one the current cluster is completed without this genotype. Another suggestion is to complete the cluster without a particular genotype if its average D^2 with the cluster is higher than the maximum among the minimum D^2 values attached to the genotypes [3]. A new cluster is tried from the remaining genotypes in a similar way. The procedure is continued until all the genotypes are exhausted.

The Tocher's method of clustering has the following disadvantages.

i) The stopping rule for formation of any cluster is arbitrary. If the suggestion from Singh and Choudhary [3] is taken for the formation of clusters, when one genotype is markedly distant from the rest, all the genotypes except this will form a single cluster.

ii) Often a genotype belonging to a cluster has on an average, a smaller D^2 value with genotypes of a different cluster than the one it belongs to.

Moreover the clustering cannot be done through a computer.

A computer oriented iterative algorithm for clustering genotypes using Mahalanobis D^2 values, which is free from the drawbacks of Tocher's method mentioned above is proposed in this paper with illustration.

METHODOLOGY

The D^2 statistic based on 'p' characteristics between any pair of genotypes was defined by Mahalanobis [1] as

$$D_p^2 = cd' W^{-1} d$$

where c—error d.f., w—matrix of mean error sum of squares and sum of products, and d'— $(X_{11}-X_{12}, X_{21}-X_{22}, \ldots\ldots, X_{p1}-X_{p2})$, X_{ij} being the mean of ith character for the jth genotype.

The D^2 values between every pair of genotypes could be determined by the method of pivotal condensation as described by Rao [4].

The iterative algorithm using D^2 values suggested herein has two parts. The first part is to form initial clusters and the second is to optimise them through iterative algorithm.

FORMATION OF CLUSTERS

The steps are summarised below.

i) Identify the two genotypes having maximum D^2 value between them as the nuclei of two clusters.

ii) Every genotype is considered in turn and allocated to the cluster for which its D^2 value with the nucleus genotype is minimum.

iii) To increase the number of clusters by one the maximum D^2 within the above two clusters is searched and the corresponding genotypes will be considered as the nuclei in addition to the nucleus genotype of the remaining cluster. The genotypes may be re-assigned as in (ii). In a similar way the number of clusters can be raised to a desired level.

ITERATIVE ALGORITHM

The clustering obtained may be optimised by the following iterative relocation algorithm.

i) Number of genotypes from 1 to v, when there are v genotypes.

ii) Take out genotype No. 1 from the cluster to which it was allotted and calculate the average intercluster D^2 value between this genotype and each cluster. (Average intercluster D^2 value between a genotype and a cluster means the arithmetic mean of the D^2 values between this genotype and each member genotype of the cluster). Allocate this genotype into that cluster for which the average intercluster D^2 value is found minimum.

iii) Repeat (ii) for all the genotypes numbered from 2 to v.

iv) With the clustering obtained in step (iii) a second iteration may be started, if necessary, i.e., repeat (ii) and (iii). The iterations have to be continued till two successive iterations end up with the same configuration of clusters.

DETERMINATION OF NUMBER OF CLUSTERS

A graphical method for determination af optimum number of clusters is suggested herein and is explained below.

A graph of weighted arithmetic mean of the average intracluster D^2 values, weights being the number of D^2 values in the cluster, against the number of clusters may be drawn. The graph will be a decreasing one. The rate of decrease also will be decreasing. The point on the X axis which is just beyond the maximum curvature could be taken as the optimum number of clusters.

ILLUSTRATION

Observations on 16 traits of 24 accessions of banana from an experiment laid out in RBD with 3 replications provided by Rajeevan [5] were utilised for illustration.

The upper triangular matrix of D^2 values between the 24 accessions, obtained by pivotal condensation method is given in Table 1.

The genotypes having maximum D^2 value are 4 and 14 and they are termed as the nuclei of two clusters. Every genotype is considered in turn and allocated that cluster for which its D^2 value with the nucleus genotype is minimum. The maximum D^2 value in these two clusters is between 1 and 14. They form the nuclei in addition to 4, the nucleus of the other cluster. Now there are three nuclei, 1, 4 and 14. All the other genotypes are allocated to these

Table 1. D^2 values between the 24 accessions of banana

Accession Nos.	2	3	4	5	6	7	8	9	10	11	12	13	14
1	16233	794	39977	42	5611	3693	20262	1285	26421	6236	6607	30035	30594
2		10121	107051	15401	2783	35310	404	8479	1513	2370	2594	2369	2316
3			51406	670	2368	7780	13487	93	18729	2802	3326	21726	22047
4				41393	75480	19549	116899	55336	131224	77752	78632	139210	140492
5					5161	4093	19246	1114	25239	5726	5984	28858	29423
6						18365	4714	1577	7873	43	327	9802	10045
7							40947	9301	49520	19434	19801	54583	55387
8								11615	593	4119	4043	1382	1240
9									16421	1951	2402	19181	19551
10										7089	6733	275	316
11											223	9004	9235
12												8770	9189
13													157
14													
15													
16													
17													
18													
19													
20													
21													
22													
23													

(Continued)

Table 1 (contd.)

Accession Nos.	15	16	17	18	19	20	21	22	23	24
1	8576	1914	24612	25548	1712	4039	5013	10052	3030	2978
2	1340	7159	1011	1213	7546	4277	3351	831	5268	5331
3	4514	316	17125	17920	215	1560	2021	5536	869	789
4	85506	59047	127267	129399	57825	69280	73088	90009	64957	64627
5	7929	1682	23512	24455	1500	3596	4638	9348	2670	2650
6	396	1059	6821	7304	1231	242	69	733	415	448
7	23340	10862	47133	48453	10329	15294	17266	25732	13342	13232
8	2581	10117	361	551	10464	6415	5538	1838	7685	7800
9	3408	79	14920	15637	69	933	1279	4332	429	392
10	4910	14557	61	83	15160	9853	8816	3929	11635	11883
11	225	1359	6116	6580	1541	320	150	489	598	634
12	232	1789	5934	643	2024	359	526	524	826	995
13	6587	17090	370	226	17924	12098	10718	5525	14071	14356
14	6849	17510	376	294	18136	12561	11054	5645	14402	14578
15		2610	4148	4546	2871	881	663	89	1446	1555
16			13139	13782	111	558	7792	3442	218	209
17				35	13714	8783	7685	3240	10413	10616
18					14435	9336	8130	3632	11043	11255
19						756	1021	3671	283	203
20							213	1436	129	234
21								1100	348	371
22									2079	2163
23										31

three clusters as before. In the same way the number of clusters can be raised to a desired level. The initial clusters thus obtained were further optimised by the iterative algorithm. The constellations of clusters for both initial and final clusterings are given in Table 2.

Table 2. Clusters obtained by the iterative algorithm using D^2 in banana

Grouping	Cluster	Genotypes in clusters	Weighted mean of intracluster D^2	No. of iterations
		Two clusters		
Initial	1	1, 2, 3, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	6264.7	
	2	4, 7		
Final	1	1, 2, 3, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	6264.7	1
	2	4, 7		
		Three clusters		
Initial	1	4	2774.9	
	2	1, 3, 5, 6, 7, 9, 11, 12, 15 19, 20, 21, 23, 14		
	3	2, 8, 10, 13, 14, 15, 17, 18, 22		
Final	1	4	2774.9	1
	2	1, 3, 5, 6, 7, 9, 11, 12, 15, 19, 20, 21, 23, 24		
	3	2, 8, 10, 13, 14, 15, 17, 18, 22		
		Four clusters		
Initial	1	4	1093.1	
	2	1, 5, 7		
	3	3, 6, 9, 11, 12, 15, 16, 19, 20, 21, 22, 23, 24		
	4	2, 8, 10, 13, 14, 17, 18		
Final	1	4	1093.1	1
	2	1, 5, 7		
	3	3, 6, 9, 11, 12, 15, 16, 19, 20, 21, 22, 23, 24		
	4	2, 8, 10, 13, 14, 17, 18		
		Five clusters		
Initial	1	4	837.4	
	2	7		
	3	2, 6, 11, 12, 15, 20, 21, 22		
	4	1, 3, 5, 9, 16, 19, 23, 24		
	5	8, 10, 13, 14, 17, 18		
Final	1	4	670.2	2
	2	7		
	3	6, 11, 12, 15, 20, 21, 22, 23, 24		
	4	1, 3, 5, 9, 16, 19		
	5	2, 8, 10, 13, 14, 17, 18		

A graph of the weighted average of intracluster D^2 values, weights being the number of D^2 values in the clusters, against the number of clusters was drawn (Fig. 1). The optimum number of clusters was determined as 4 where the curve has the maximum curvature.

The 24 genotypes were also grouped by Tocher's method. There are five clusters by this method. The cluster configurations along with the average intra cluster D^2 values are given in Table 3.

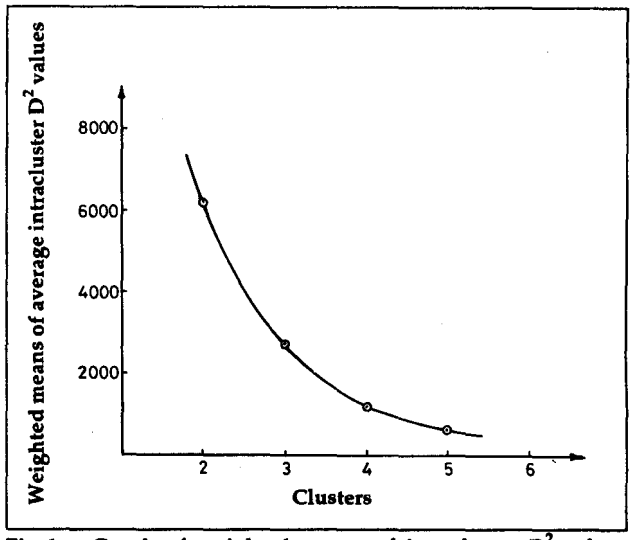

Fig. 1. Graph of weighted means of intracluster D^2 values against number of clusters

Table 3. Clusters obtained by Tocher's method

Cluster	Genotypes in clusters	Weighted mean of intracluster D^2
1	3, 6, 9, 11, 12, 15, 16, 19, 20, 21, 22, 23, 24	1036.4
2	2, 8, 10, 13, 14, 17, 18	
3	1, 5	
4	4	
5	7	

It may be noted that in the case of clustering by iterative algorithm the weighted average of the intra cluster D^2 values was 670.23 for five clusters against 1036.41 in the case of Tocher's method. This shows the superiority of the new method over Tocher's method in achieving homogenity of genotypes within clusters.

DISCUSSION

The iterative algorithm proposed herein achieves a clustering of genotypes free from the drawbacks of the Tocher's method. Every genotype is allocated to that cluster for which it is more homogenious, which is the basic principle of any clustering procedure. This is evidenced by the very low value of the weighted arithmetic mean of intra cluster D^2 values compared to Tocher's method in the illustration. More over clustering by the procedure suggested herein can be done in a computer while that by Tocher's method cannot be.

A FORTRAN programme for the clustering by the method proposed herein is given in Appendix 1.

APPENDIX 1

```
      PROGRAM CLST
C     Programme to group genotypes by the iterative relocation algorithm
C     Based on Mahalanobis D2 values
      DIMENSION A (50, 50), KS (10, 50), G (3), KN (50), KK (10)
C     Inputs
C     N—Number of genotypes
C     KZ—Maximum number of clusters into which they are to be grouped
C     G—The name of the file containing N x N matrix of D2 values
C     II—The drive number having the disk containing the data file
C     Output will be the cluster configurations for initial as well as final solutions
      corresponding to two to KZ clusters and the Corresponding average
      intracluster D2 values
      READ (1, 50), G, I1
      READ (1, 51) N, KZ
      CALL OPEN (6, G, I1)
      Do 90 I = 1, N
90    READ (6, 52) (A (I, J), J = 1, N)
52    FORMAT (6E15.8)
50    FORMAT (2A4, A3, I1)
51    FORMAT (2I2)
      KK (1) = N
      Do 1 I = 1, N
1     KS (1, I) = I
      K = 1
100   A1 = 0
      Do 2 I=1, K
      If (KK(I).Eq.1) go to 2
      KL = KK (I)-1
```

(Continued)

```
      KL1=KK(I)
      Do 2 J=1, KL
      J1 = J+1
      Do 2 JJ=J1, KL1
      K1=KS (I, J)
      K2=KS (I, JJ)
      If (A1.GT.A(K1, K2)) go To 2
      A1=A (K1, K2)
      KM=K1
      K0=K2
      KI=I
2     CONTINUE
      K=K+1
      KS(KI, 1)=KM
      KS(K, 1)=K0
      Do 3 I=1, K
      KI=KS (I, 1)
      KN(KI)=I
3     KK(I)=1
      Do 6 I=1, N
      Do 4 L=1, K
      If (I.EQ.KS (L, 1)) go t0 6
4     CONTINUE
      L1=KS(1, 1)
      A1=A (I, L1)
      LK=1
      Do 5 L=2, K
      L1=KS (L, 1)
      If (A1.LT.A (I, L1)) go to 5
      A1=A (I, L1)
      LK=L
5     CONTINUE
      KK(LK)=KK(LK)+1
      KM=KK(LK)
      KS(LK, KM)=I
      KN(I)=LK
6     CONTINUE
      Write(2, 95) K
```

(Continued)

```
95      FORMAT (10X, 'Number of Clusters', I5/)
        Call CLUST (A, KN, N, K)
        If (K.NE.KZ) go to 100
        STOP
        END
        SUBROUTINE CLUST (X, M, NV1, NK)
        DIMENSION M (50), M0(10), M1(10), MA(10, 50), MB(10,50)
        DIMENSION X(50, 50), G(3), Y(10, 10)
        COMMON XL
        Do 93 I=1, NK

93      M0(I)=0

17      FORMAT (2A4, A3, I1)

10      FORMAT (20I2)

15      FORMAT (6E15.8)
        Do 1 I=1,NV1
        MI=M(I)
        M0(MI)=M0(MI)+1
        ML=M0(MI)

1       MA(MI, ML)=I
        Do 2 I=1, NK
        M1(I)=M0(I)
        MI=M1(I)
        Do 2 J=1, MI

2       MB(I, J)=MA (I, J)
        Call BET (X, M1, MB, Y, NK)
        KK=1

500     Do 25 I=1, NV1
        LN=M(I)
        If (M1(LN).LE.1) go to 25
        MI=M(I)
        MK=M1(MI)
        M1(MI)=M1(MI)-1
        MK1=MK-1
        LN=1
        M(I)=1
        Do 20 J=1, MK1
        If (MB(MI, J).EQ.I) go to 200
```

(Continued)

```
20      CONTINUE
        Go to 202

200     Do 201 K=J, MK1

201     MB(MI, K)=MB(MI, K+1)

202     Do 21 L=1, NK
        MI=M1 (L)
        DX=0
        Do 50 K1=1, MI
        K2=MB (L, K1)

50      DX=DX+X (I, K2)
        DX=DX/M1 (L)
        If (L.EQ.1) DA=DX
        DB=DX
        If (DB.GE.DA) go to 21
        LN=L
        M(I)=L
        DA=DB

21      CONTINUE
        M1(LN)=M1 (LN) + 1
        MI=M1 (LN)
        MB(LN, MI)=I

25      CONTINUE
        Do 250   I = 1,  NK
        If (M1(I).N.E.M0(I) go to 252

250     CONTINUE
        Do  251 I=1, NK
        MI=M1(I)
        Do 251 J=1, MI
        If (MB(I, J).NE.MA(I, J)) go to 252

251     CONTINUE
        Go to 254
```

(Continued)

```
252     KK=KK+1
        Do 253 I=1, NK
        M0(I) = M1(I)
        Do 253 J=1, MI
        MA(I, J)=MB(I, J)

253     CONTINUE
        Call BET (X, M1, MB, Y, NK)
        Go to 500

254     WRITE (2, 300) KK

300     FORMAT (4X, 'No. of iteration=', I4)

101     RETURN
        END
        Subroutine BET (X, M1, MB, Y, NK)
        DIMENSION Y(10, 10), M1(10), MB(10, 50), X(50, 50)
        Do 1 I=1, NK
        Do 1 J=I, NK

1       Y(I, J) = 0
        NK1=NK-1
        Do 2 I=1, NK1
        MI=M1 (I)
        If (M1(I).LE.1) g0 to 4
        MI1=MI-1
        Do 3 K=1, MI1
        K0=MB (I, K)
        K1=K+1
        Do 3 K2=K1, MI
        K3=MB (I, K2)

3       Y(I, I)=Y(I, I)+X(K0, K3)
```

ACKNOWLEDGMENTS

The authors are grateful to Dr. P. K. Rajeevan, College of Horticulture, Trichur for providing the data, and to the referee for suggestions to improve.

REFERENCES

1. P. C. Mahalanobis. 1936. On the generalized distance in statistics. Proc. Natl. Sci. Acad. India., **2**: 49–55.

2. C. R. Rao. 1948. The utilisation of multiple measurements in problems of biological classification. J. Roy. Stat. Soc. B, **10**: 159–203.

3. R. K. Singh and B. D. Choudhary. 1979. Biometrical Methods in Quantitative Genetics Analysis. Kalyani Publishers, New Delhi.

4. C. R. Rao. 1952. Advanced Statistical Methods in Biometric Research. John Wiley and Sons, Inc., New York.

5. P. K. Rajeevan. 1985. Intraclonal Variations and Nutritional Studies in Banana cv. Palayankodan. Ph. D. Thesis. Kerala Agricultural University, Trichur.