A new split-and-merge clustering technique

D. Chaudhuri, B.B. Chaudhuri and C.A. Murthy

Electronics & Communication Sciences Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India

Received 5 August 1991

Abstract

Chaudhuri, D., B.B. Chaudhuri and C.A. Murthy, A new split-and-merge clustering technique, Pattern Recognition Letters 13 (1992) 399–409.

A new clustering algorithm is developed for efficient classification of data in $\mathbb{R}^q$ when there exists no a priori information about the number of clusters. The algorithm is based on a split-and-merge technique. The type-I splitting is guided by density of data over strips at different directions around the centroid of the data. The type-II splitting is the usual K-means clustering algorithm ($K=2$) and rechecked with the help of a merging technique. A theorem on the convergence of this algorithm is proved.

Keywords. Cluster analysis, split-and-merge, K-means clustering.

1. Introduction

Clustering is a useful and important technique in image processing and pattern recognition [1,2,5, 7,9]. There exist two classes of clustering techniques, namely hierarchical and non-hierarchical techniques. Among non-hierarchical techniques K-means and ISODATA are popular. Of them, ISODATA is a split-and-merge technique of achieving a prespecified number of clusters. Among other split-and-merge techniques Wishart [4] as well as Liu and Tsai [3] may be mentioned. The method in this paper also falls in this category.

The main difference of the proposed method with those of the others is that here we try to split the clusters by noting the density at different directions by observing the data over strips. In order to overcome some defects of this approach, another splitting approach, that is, the simple 2-means algorithm under a certain restriction is used. Also, for merging, it is tested whether the data at the boundary of a cluster is very close to the data at the boundary of the other cluster. The splitting and merging techniques are described in Section 2. Section 3 describes the proposed algorithm with the corresponding flowchart and the theorem of convergence criterion. The results on synthetic and real data (remotely sensed imagery) are presented in Section 4.

Correspondence to: B.B. Chaudhuri, Electronics & Communication Sciences Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700035, India.

2. Splitting techniques

In type-I splitting, strips of finite width at different directions around the centroid of the data are considered. The data is split across the *sparsely populated strip*. For a two-dimensional data set we consider four directions at the center of the data as

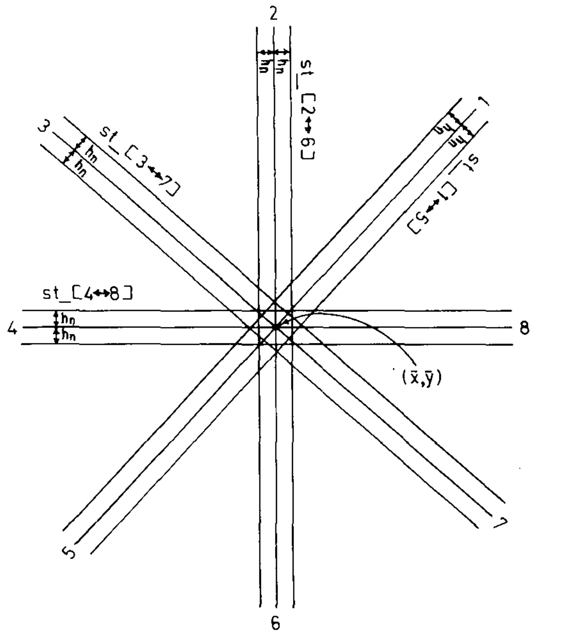

Figure 1. Strips in the four directions in a 2-dimensional data set.

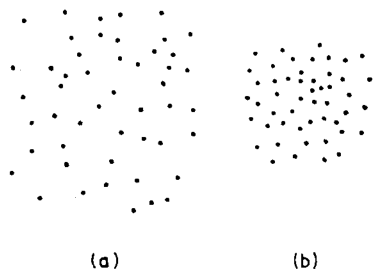

Figure 2. Clusters with different density. (a) Cluster of 50 patterns with relatively large interpoint distances. (b) Cluster of 50 patterns with relatively small interpoint distances.

shown in Figure 1. They are named as $1 \leftrightarrow 5$ (one diagonal), $2 \leftrightarrow 6$ (vertical), $3 \leftrightarrow 7$ (another diagonal), and $4 \leftrightarrow 8$ (horizontal). The corresponding strips are denoted by St_$[1 \leftrightarrow 5]$, St_$[2 \leftrightarrow 6]$, St_$[3 \leftrightarrow 7]$ and St_$[4 \leftrightarrow 8]$, respectively. For a three-dimensional data set we consider $2^{3-1}+3=7$ directions at the centroid of the data ($2^{3-1}=4$ directions are 4 diagonals and 3 directions are 3 axes). The number of directions considered for a q-dimensional data set is $2^{q-1}+q$. The strips are constructed along (i) the q axes and (ii) the diagonals (note that the number of diagonals in q dimensions is 2^{q-1}). But, observe that the greater the number of directions for strips the better the accuracy of the result is.

The width of a strip along any direction is to be found out before actually constructing the strip. The width is an important impediment in deciding whether the data is indeed sparsely populated in that direction. If the width is very large then all the points in the data set may belong to the strip. If it is very small then the strip may not be amenable for making any decision. In this connection a measure of finding the width is described below using an example.

Example. Three data sets are shown in Figures 2(a), 2(b) and 3(a) where the number of points in these data sets are the same. Intuitively, the set of points in Figure 2(b) is a single cluster. The data in Figure 2(a) can be called a single cluster though interpoint distances are generally greater than in Figure 2(b). In Figure 3(a) it is intuitively clear that there are two clusters. Thus the sets of points in Figures 2(a) and 2(b) should not be split while those in Figure 3(a) are to be split. A way of splitting the set of points in Figure 3(a) at the center (X_0) is shown in Figure 3(b).

Draw a strip of width $2h_n$ as shown in Figure 3(b). Count the number of points in the strip. If the number of points is less than some threshold, say θ, then conclude that there are two clusters. Note that, if this procedure is to be followed in Figures 2(a) and 2(b) then h_n should be taken suitably. h_n should not be too big so that it could

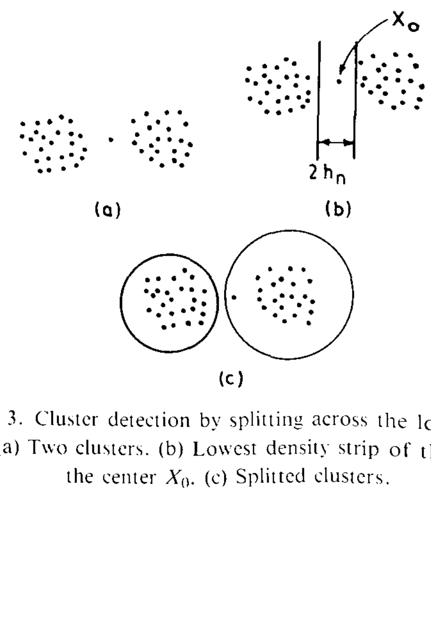

Figure 3. Cluster detection by splitting across the low-density strip. (a) Two clusters. (b) Lowest density strip of the data at the center X_0. (c) Splitted clusters.

include the entire set of points in the strip. h_n should not be too small so that the set of points in Figure 2(a) may be split. Note that h_n in a sense gives the connectivity of the set. In this regard a result [20] is stated in the Appendix. Based on this result we propose that the width of the strip should be $2h_n$ where $h_n = a\varepsilon_n, \varepsilon_n = 1/n^p, 0<p<1/q$ and a is a constant, n is the total number of points in the q-dimensional data set.

The type-I splitting is described below for $\mathbb{R}^2$.

Let the given set of points be

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^2.$$

Let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Type-I splitting is applied at the mean point of the data set [i.e., at $(\bar{x}, \bar{y})$] along a particular direction among the four directions discussed before.

Let b_1, b_2, b_3 and b_4 be the number of points in each of the strips St_[$1 \leftrightarrow 5$], St_[$2 \leftrightarrow 6$], St_[$3 \leftrightarrow 7$] and St_[$4 \leftrightarrow 8$], respectively.

Let $n_1 = \min\{b_1, b_2, b_3, b_4\}$.

Find the strip in which the number of points is n_1. If n_1 is not very small compared to n, then there should not be any split in that direction. If

$$\frac{n_1}{n} \times 100 < l_1 \tag{1}$$

then split the set S along the corresponding direction. Inequality (1) is called the *splitting restriction*. Here l_1 is a small quantity dependent on the width of the strip. If the number of points in a strip satisfies (1) then the strip is called a *sparsely populated strip*. Otherwise the strip is called a *densely populated strip*.

Note that if in two or more directions, the number of points, n_1, satisfies the *splitting restriction*, then one of the directions is chosen arbitrarily for splitting.

Type-I splitting can also be extended to three or more dimensions. Observe that for a two-dimensional data set, two straight lines are needed for constructing a strip (Figure 1). For higher-dimensional data sets, hyperdimensional strips need to be constructed. The number of hyperplanes needed for constructing a hyperdimensional strip in q dimensions is 2^{q-1}. The inequality (1) would remain unchanged for q-dimensional data.

Note that *sparsely populated regions* need not always be present at the center of the data. They can occur at other places as well. In this context a splitting method, namely type-II splitting is incorporated in the algorithm.

The type-II splitting technique is the usual K-means method by Forgy [6] with some restriction imposed and $K = 2$. The selection of initial seed points is to be done suitably [14].

After the 2-means algorithm converges, let m_1 and m_2 be the mean points of the two subclusters S_1 and S_2 of S.

Let $d_0 = \|m_1 - m_2\|$.

The *almost equidistant point set*, A, is the collection of those points whose differences of the distances from m_1 and m_2 are less than 10% of d_0, i.e.,

$$A = \left\{ z: \quad z \in S, \left| \|z - m_1\| - \|z - m_2\| \right| < \frac{d_0}{10} \right\}$$

and $n_4 = \# A$. If

$$\frac{n_4}{n} \times 100 < l_2 \tag{2}$$

then S_1 and S_2 are said to be two different clusters. The inequality (2) is called the *almost equidistant restriction*. As before, l_2 is a small quantity.

If $(n_4/n) \times 100 \geqslant l_2$ then the merging criterion will have to be checked for deciding whether subclusters S_1 and S_2 remain divided or not.

Merging technique

Let $S = \{z_1, z_2, \ldots, z_n\} \subseteq \mathbb{R}^q$. Let S_1 and S_2 be such that $S_1 \cap S_2 = 0$, $S_1 \cup S_2 = S$ and they are the outcome of the 2-means algorithm on S. Let

$$m_i = \left(\sum_{z \in S_i} z \right) \Big/ \# S_i, \quad i = 1, 2.$$

Let

$$A = \left\{ z: \quad z \in S, \left| \|z - m_1\| - \|z - m_2\| \right| < \frac{d_0}{10} \right\}$$

where $d_0 = \|m_1 - m_2\|$. Let $n_4 = \# A$. Let

$$m_3 = \frac{1}{n_4}\left(\sum_{z \in A} z \right)$$

and

$$H = \left\{ z: \ \|m_3 - z\| \leqslant \frac{d_0}{5} \right\}.$$

The set H is called the *merging circle set* (Figure 4). Let

$$n_5 = \# H \cap S_1 \quad \text{and} \quad n_6 = \# H \cap S_2.$$

If

$$\frac{|n_5 - n_6|}{n} \times 100 < l_3 \tag{3}$$

then only S_1 and S_2 are to be merged. l_3 must be very small. The inequality (3) is called the *merging restriction*.

It may be noted that

1. The above method basically verifies whether the points near the boundary of the two clusters have equal representation in both clusters (Figure 4).
2. The above method gives a criterion for merging $S_1 \cup S_2$ which are products of the 2-mean algorithm on S.
3. If any one of S_1 and S_2, say S_1, is further subdivided (say to S_{11} and S_{12}), then also the same merging technique can be applied. The definitions m_i, A, d_0, n_4, m_3, H, n_5 and n_6 are unchanged.

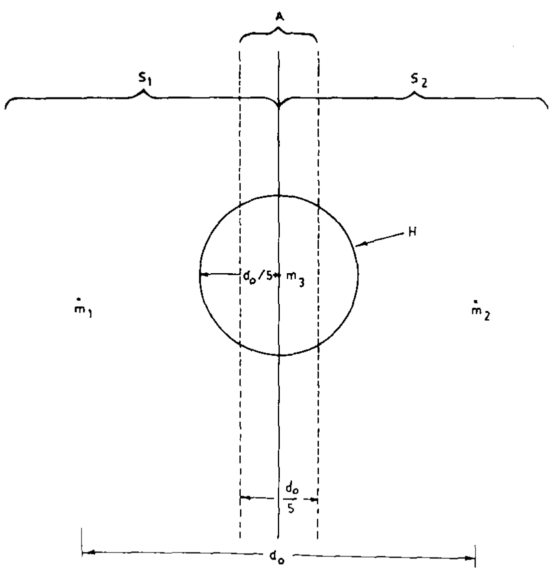

Figure 4. The process of merging in 2-dimension.

Let S_{11} be such that $A \cap S_{11} \neq \emptyset$. It is then generally true that $A \cap S_{12} = \emptyset$. Thus S_{11} and S_2 are to be merged if the merging criterion is satisfied.

4. If both S_1 and S_2 are further subdivided then the merging technique is to be applied similarly.

3. Proposed algorithm

In the proposed algorithm the number of iterations is represented by J while the number of clusters at the Jth iteration is given by K_J. Initially

$$J = 0 \quad \text{and} \quad K_0 = 1.$$

Thus, there exists one cluster initially which is S. The algorithm has the following steps.

Step 1. Apply type-I splitting on every cluster.

Step 2. If no cluster is divided in Step 1, go to Step 3. Otherwise go to Step 1.

Step 3. Apply the 2-means algorithm on every cluster. Find out the *almost equidistant point sets* for every cluster. Check the inequality (2) for every one of the *almost equidistant point sets* which do not satisfy the *almost equidistant restriction*; then go to Step 4. Otherwise go to Step 5.

Step 4. Let $A_1, A_2, \ldots, A_L$ be those *almost equidistant point sets* which do not satisfy the *almost equidistant restriction*. Let the corresponding clusters for $A_1, A_2, \ldots, A_L$ be $C_1, C_2, \ldots, C_L$. Let every C_s be divided into C_{s1} and C_{s2} in Step 3 for $s = 1, 2, \ldots, L$. Apply type-I splitting on C_{s1} and C_{s2} for $s = 1, 2, \ldots, L$. For every $s = 1, 2, \ldots, L$, the following two cases arise.
(a) None of C_{s1} and C_{s2} is split. Then check the *merging restriction* on C_{s1} and C_{s2} using A_s.
(b) At least one of the C_{s1} and C_{s2} is split. Then check the *merging restriction* for those sets with which A_s has non-empty intersection.

Step 5. $J \leftarrow J + 1$. Find the number of cluster K_J.

Step 6. If $K_J = K_{J-1}$ then go to Step 7. Otherwise go to Step 1.

Step 7. Stop.

The flowchart of the above algorithm is shown in Figure 5.

Note that the algorithm converges when the

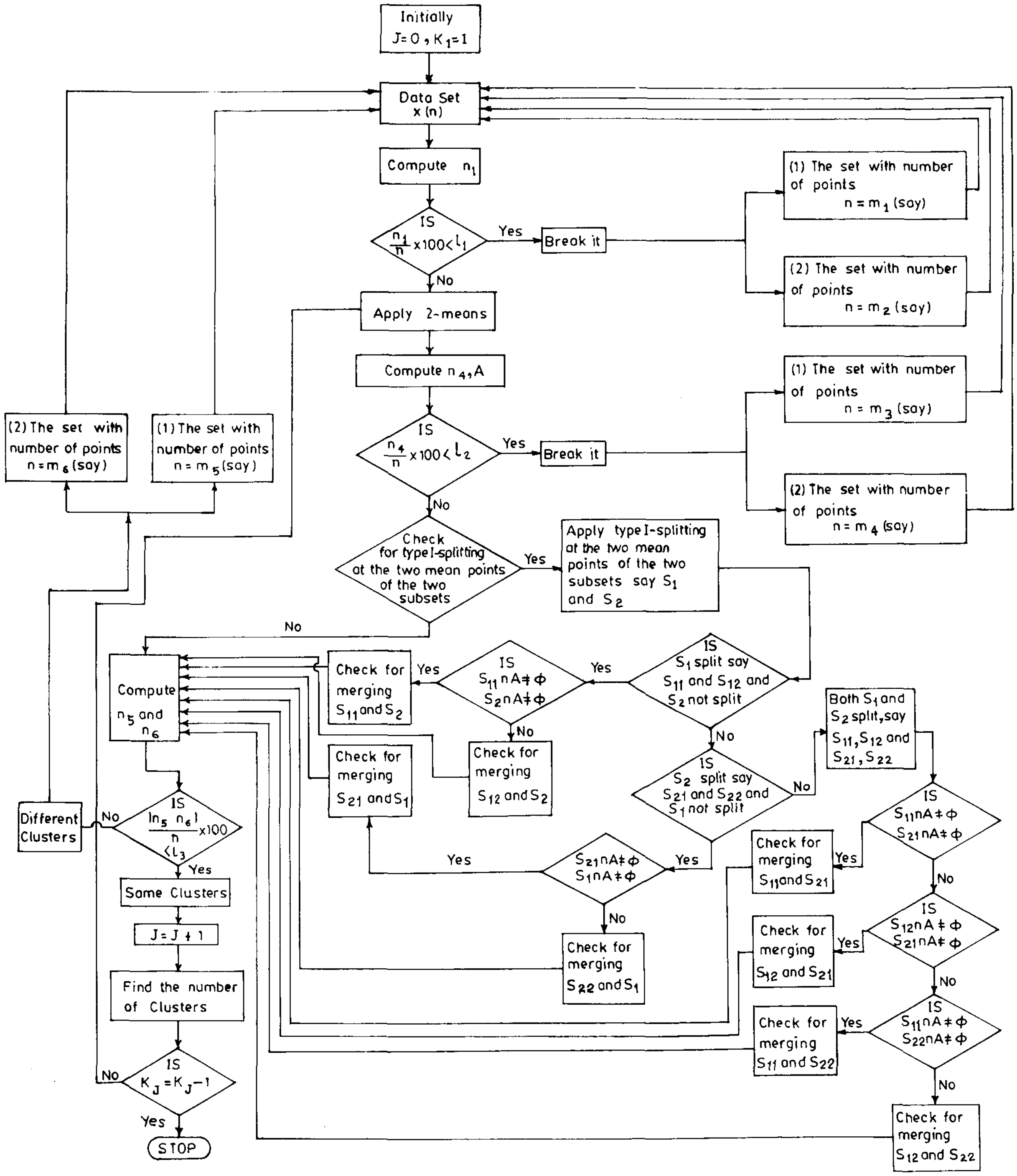

Figure 5. A flowchart of the proposed algorithm.

number of clusters in the Jth iteration and the $(J+1)$th iteration are equal. A theorem is stated below which shows that if the number of clusters in the Jth and $(J+1)$th iterations are equal then the clusters at the end of Jth iteration are the same as the clusters at the end of the $(J+1)$th iteration. Theorem 1 and Theorem 2 stated below are proved in the Appendix.

Theorem 1. *Let the number of clusters at the end of the Jth and $(J+1)$th iterations be k and l respectively. Then $k \leqslant l$.*

Theorem 2. *Let the clusters at the end of the Jth and $(J+1)$th iterations be $P_1, P_2, \ldots, P_k$ and $Q_1, Q_2, \ldots, Q_k$. Then*

$$\{P_1, P_2, \ldots, P_k\} = \{Q_1, Q_2, \ldots Q_k\}.$$

4. Experimental results

4.1. On artificial data

The clustering scheme described in the paper has been implemented on various data sets. The programs are run on an IBM PC/AT microcomputer in TURBO PASCAL language. In all the tested cases, the breadth of the strip, i.e., $2h_n = 2a\varepsilon_n$ [$\varepsilon_n = 1/n^p$, $0<p<0.5$ and $a=3$, $p=0.05$ and n = number of points], the values of l_1, l_2 and l_3 are taken 4, 5, 3 and 0.7, respectively.

Figure 6(a) shows multi-cluster data of size 204. For type-I splitting of this data, the mean point is marked by 'X', the direction is marked by a dotted line (3 ---- 7) shown in Figure 6(b). Here $n_1 = 21$ and also St_[1 ↔ 5] and St_[3 ↔ 7] contain 21 points each. But

$$\frac{n_1}{n} \times 100 \approx 10.3 > 5$$

so the strip is *densely populated* and is not to be split. The results of type-II splitting are shown in Figure 6(c). The mean points of the two subsets S_1 and S_2 (obtained by using the 2-means algorithm) are marked by 'X_1' and 'X_2' and the boundary line is marked by 'B_1B_2'. The *almost equidistant points* are marked by '$\odot$'. Here $n_4 = 15$. The mean point of the *almost equidistant points* is marked by 'C'.

Since $(n_4/n) \times 100 = 7.3 > 3$, the merging criterion is to be checked. For S_1 and S_2, the type-I splitting technique is also to be verified. Thus for S_1, 4 strips are generated with width 4 at the center X_1 and the minimum size of the strip is found for strip St_[2 ↔ 6] with size 9. Since $(9/n) \times 100 = 4.4 < 5$, S_1 should indeed be split at the center along the direction $2 \leftrightarrow 6$. Similarly it can be seen that S_2 should not be split at the center X_2.

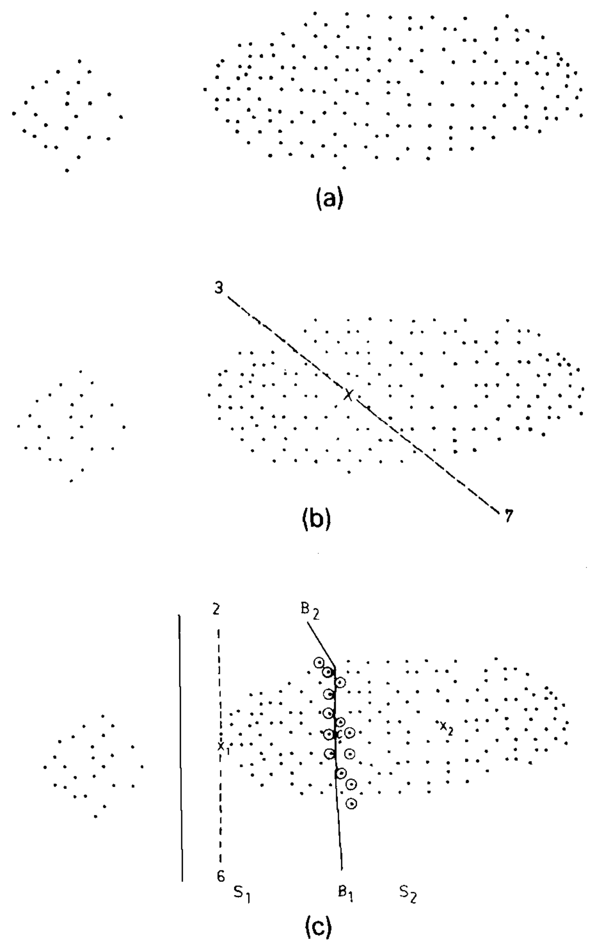

Figure 6. Synthetic data to show the need of type-II splitting. (a) A multi-cluster data of size 204. (b) The mean point and the direction of the low-density strip. (c) The resulting clusters.

The equidistant point set, H, for the pair S_1 and S_2 is found. The number of points within the *merging circle* is found to be 25, the values for d_0, n_5 and n_6 are found to be 22.3, 13 and 12, respectively. Here

$$\frac{|n_5 - n_6|}{n} \times 100 = 0.4 < 0.7.$$

So one of the divided portions of S_1 is to be merged with S_2. Thus at the end of the first iteration, the number of clusters is found to be 2 (Figure 6(c)).

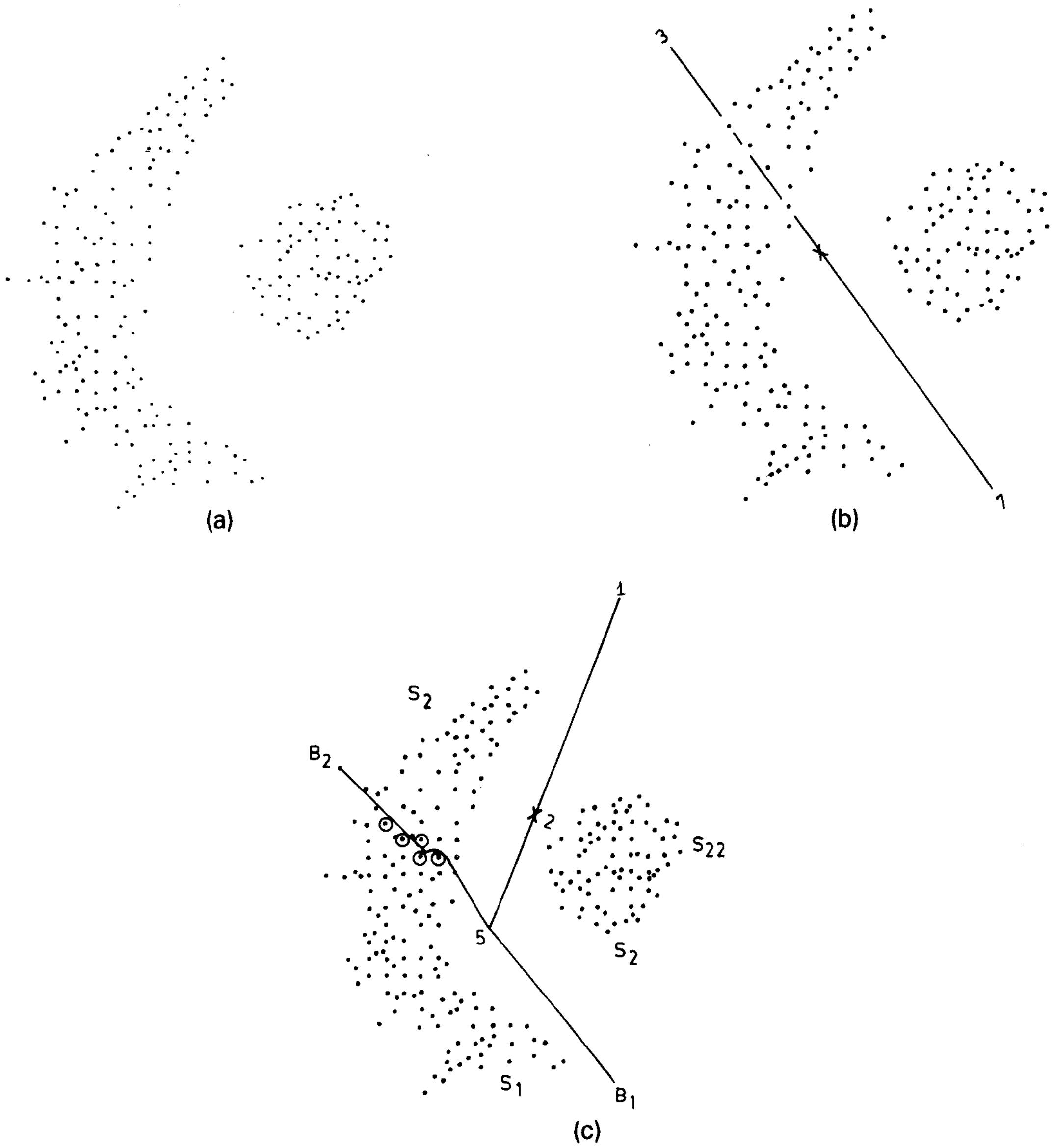

Figure 7. Another example. (a) A multi-cluster data of size 237. (b) The mean point and the direction of the low-density strip. (c) The resulting clusters.

On each one of the above two clusters, type-I splitting and type-II splitting are applied. It has been found that no further divisions can take place. Thus at the end of the second iteration too, the number of clusters is the same. So the process is terminated. One cluster has 25 points and the other has 179 points.

The results of the algorithm on other data sets are also demonstrated. Figure 7(a) shows a data set and the corresponding intermediary step is shown in Figure 7(b) where the mean point is marked by 'X', the direction is marked by a dotted line (3 ---- 7). But the strip along (3 ---- 7) is *densely populated* and is not to be split. The results of type-II splitting are shown in Figure 7(c). The boundary line of the two subsets S_1 and S_2 is marked by 'B_1B_2'. For S_2, the minimum size of the strip is found for strip $St_-[1 \leftrightarrow 5]$ in the direction $1 \leftrightarrow 5$ at the center 'X_2'. The minimum number of points in St_$[1 \leftrightarrow 5]$ satisfies the *splitting restriction*. S_2 should indeed be split at the center 'X_2' along the direction $1 \leftrightarrow 5$. S_{21} and S_1 are merged. Hence the two clusters are S_{22} and $S_{21} \cup S_1$ which are shown in Figure 7(c).

Figure 8(a) shows another data set which can be split by type-I splitting only and the coresponding results are shown in Figure 8(b). No merging is needed in this case.

The results of applying the proposed algorithm on a real-life data are given below.

4.2. *On remotely sensed data*

Analysis of satellite imager has wide applications such as crop yield estimation, estimation of forest regions etc. [10,11]. Satellite images are used for defence applications too. The proposed algorithm was applied on an Indian remote sensing satellite (IRS) image.

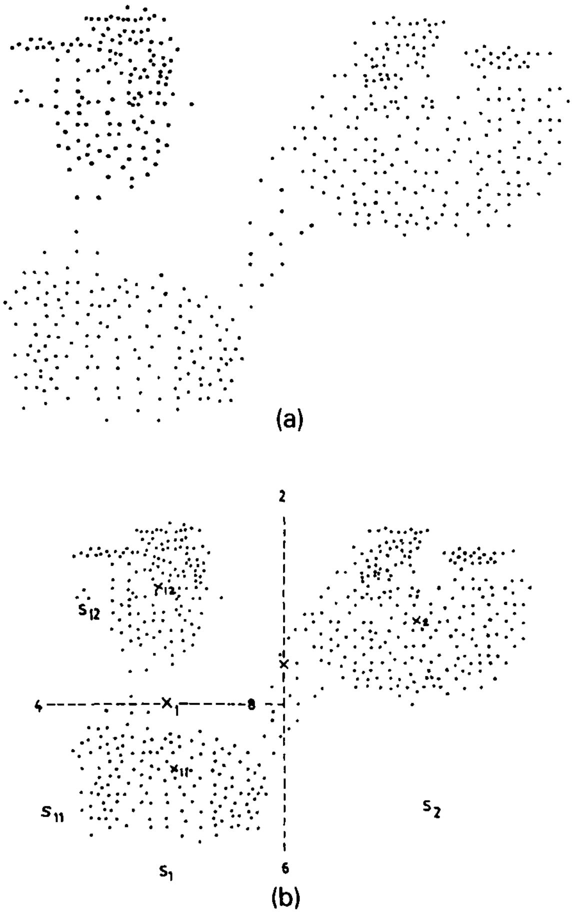

Figure 8. An example where type-I splitting only is needed. (a) The data set. (b) Resulting clusters.

Table 1

Band	Wavelength
Blue	0.45 μm–0.52 μm
Green	0.52 μm–0.59 μm
Red	0.62 μm–0.68 μm
Infrared	0.77 μm–0.86 μm

The IRS provides images of two ground resolutions, namely 72.5 m × 72.5 m and 32.5 m × 32.5 m. It has four bands namely blue, green, red and infrared. The wavelengths of these bands are given in Table 1 [12].

The proposed algorithm has been applied on an IRS image of ground resolution 36.25 m × 36.25 m. The area under consideration in that image is a suburb near Calcutta namely Barrackpore. The images corresponding to green and infrared bands for that scene are given in Figures 9 and 10, respectively. The observed gray value in the infrared band has the range 16 to 66 and the green band has the range 18 to 51. The bivariate frequency table of the gray values of these two bands is given in Table 2. The clustering partition by the proposed algorithm is shown as the dashed line in Table 2. The clusters mapped back in the image space are shown in Figure 11. Here, the white pixels in

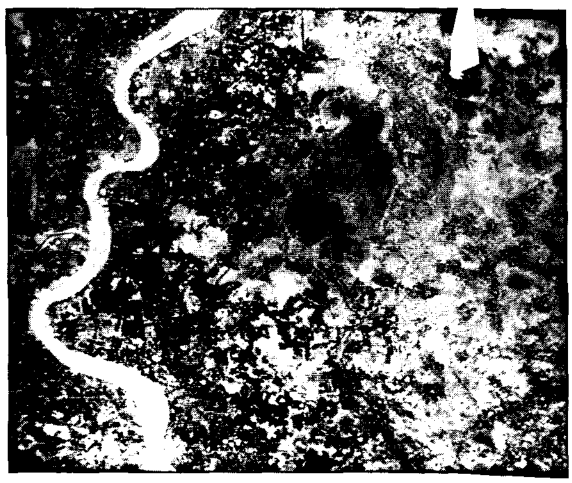

Figure 9. Remotely sensed image corresponding to green band.

Figure 10. Remotely sensed image corresponding to infrared band.

Figure 11. Resulting two clusters mapped in image domain.

Figure 11 give the water pixels in the scene. The Hooghly river in the scene is demarkated distinctly from the rest. The path of the river is clearly seen in Figure 11.

Acknowledgement

The authors wish to thank Prof. D. Dutta Majumder for his interest in the work. The authors also acknowledge Mr. J. Gupta for typing the manuscript and Mr. S. Chakraborti for his drawings.

Appendix

Result 1 [13]. Let $\varepsilon_n \to 0$ and $n\varepsilon_n^q \to \infty$, $\varepsilon_n > 0$ $\forall n$ and q is a positive integer $\geqslant 2$.

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random vectors following uniform distribution on α, where $\alpha \subseteq \mathbb{R}^q$, α is unknown and $\lambda(\delta\alpha) = 0$ [λ is the Lebesgue measure in q dimensions and $\delta\alpha$ is the boundary of α]. Let

$$\alpha_n = \bigcup_{i=1}^{n} \{x \in \mathbb{R}^q: \quad \|x - x_i\| \leqslant \varepsilon_n\}.$$

Then α_n is a consistent estimator of α, i.e.,

$$E_\alpha[\lambda(\alpha_n \triangle \alpha)] \to 0 \quad \text{as } n \to \infty.$$

[E represents expectation and $\triangle$ represents symmetric difference.]

If this estimation procedure for a suitably selected ε_n is applied to the data in Figure 2(a), then the approximate output is shown in Figure 2(c).

Let $S \subseteq \mathbb{R}^q$ and $G = \{x_1, x_2, \dots, x_n\} \subseteq S$. Now let

$$d_x = \inf_{\substack{y \in G \\ y \neq x}} d(x, y) \quad \forall x$$

where $d(x, y)$ is the distance between two points x and y. Let

$$b = \max_{x \in G} d_x.$$

Then note that if $b > 2\varepsilon_n$ then the estimated set of G will have at least two components. Thus, G will be partitioned.

Hence, if the data is to be checked for splitting at X_0 of Figure 2(a) then note that h_n should be greater than or equal to ε_n. The choice of h_n will be $h_n = a\varepsilon_n$ where $\varepsilon_n = 1/n^p$, $0 < p < 1/q$ and a is a constant.

Proof of Theorem 1. Let the clusters at the end of the Jth and $(J+1)$th iterations be $P_1, P_2, \dots, P_k$ and $Q_1, Q_2, \dots, Q_l$ respectively.

It is known that there are k clusters in the beginning of the $(J+1)$th iteration. Each iteration treats the clusters present at the beginning independently.

Table 2
Bivariate frequency table of gray levels for green and infrared band images

← Gray value in green band →

← Gray value in infrared band →	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
16	1	10	15	31	7	2	3	3	0	0	0	0	0	0	0
17	0	2	25	16	5	4	4	12	1	4	0	1	0	0	0
18	0	4	7	5	4	3	8	6	4	5	3	0	0	0	0
19	5	8	13	16	10	17	11	8	5	12	6	3	5	5	11
20	2	8	21	37	9	25	35	20	18	12	5	10	49	80	24
21	2	7	28	34	37	36	58	23	19	19	13	278	1760	1428	51
22	0	5	19	53	41	68	107	54	25	18	23	232	1831	1874	39
23	1	6	26	60	84	79	126	61	43	19	37	87	407	308	16
24	0	4	33	78	122	106	129	54	25	38	38	52	116	79	10
25	0	2	32	97	169	143	159	79	69	49	46	37	91	60	5
26	0	2	51	139	224	237	251	110	63	54	45	36	75	48	7
27	0	2	26	160	277	339	272	216	143	69	50	44	59	36	6
28	0	2	32	181	312	484	589	416	323	129	79	55	48	29	9
29	0	1	36	206	396	647	953	920	794	342	104	57	60	23	6
30	0	4	34	237	578	958	1503	1593	1505	759	222	93	56	38	2
31	0	2	35	298	781	1388	2135	2354	2348	1372	470	140	73	27	8
32	0	1	23	391	1089	1854	2657	2923	2890	1640	706	233	94	47	11
33	0	1	33	506	1474	2453	3482	3470	3183	1909	891	295	128	44	30
34	0	0	22	439	1669	2862	3780	3596	3386	2114	1039	365	125	52	31
35	0	0	13	310	1500	2928	3910	3728	3375	2293	1259	882	183	82	48
36	0	0	9	188	1295	2729	3902	3744	3396	2597	1572	692	254	95	48
37	0	0	6	150	1073	2635	3772	3862	3784	2739	1902	954	385	141	50
38	0	0	0	69	789	2214	3520	3863	3663	2932	2297	1325	560	209	183
39	0	0	2	45	524	1721	2755	3162	3510	2967	2491	1716	702	294	126
40	0	0	1	27	306	1142	1931	2425	2833	2735	2277	1405	808	305	125
41	0	0	1	20	209	743	1307	1588	2069	2181	1899	1467	852	396	201
42	0	0	0	12	144	457	796	1151	1509	1561	1527	1140	728	375	196
43	0	0	0	11	91	265	496	744	1052	1106	1080	865	5. 3	304	161
44	0	0	0	5	62	177	329	516	685	778	713	594	352	223	148
45	0	0	0	11	66	108	238	355	527	542	482	339	214	139	92
46	0	0	1	16	45	83	185	258	365	97	340	225	155	82	55
47	0	0	0	14	29	66	127	216	272	321	201	145	89	68	39
48	0	0	0	10	36	36	104	162	265	255	190	106	71	42	35
49	0	0	0	3	27	32	68	141	219	190	122	79	58	36	35
50	0	0	0	13	15	18	48	122	181	135	84	55	36	30	21
51	0	0	1	4	16	13	43	99	137	124	60	24	18	16	12
52	0	0	0	3	7	16	26	83	131	96	41	19	8	9	4
53	0	0	0	1	6	8	26	55	115	71	25	11	3	5	8
54	0	0	0	0	4	6	14	50	85	38	17	9	3	2	2
55	0	0	0	0	1	8	18	46	58	34	16	5	0	2	2
56	0	0	0	2	3	4	20	19	39	25	12	3	0	0	0
57	0	0	0	0	0	1	5	28	36	18	2	2	1	0	0
58	0	0	0	0	0	3	4	14	22	7	6	0	0	0	0
59	0	0	0	0	0	2	8	13	11	12	2	1	0	0	0
60	0	0	0	0	0	1	2	12	6	3	0	0	1	0	0
61	0	0	0	0	0	3	3	4	3	1	2	0	0	0	0
62	0	0	0	0	0	1	3	4	2	1	0	1	0	0	0
63	0	0	0	0	1	1	3	3	4	1	0	1	0	0	0
64	0	0	0	1	0	0	1	3	0	0	0	0	0	0	0
65	0	0	0	0	0	2	1	4	0	0	0	0	1	0	0
66	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0

← Gray value in infrared band →	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	2	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	6	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	10	5	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	14	5	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	18	11	6	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	34	16	6	6	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0
38	47	38	11	8	3	4	1	0	0	0	1	0	0	0	0	0	0	0	0
39	50	32	25	18	3	1	1	0	1	0	0	0	0	0	0	0	0	0	0
40	63	49	31	20	5	1	4	1	0	0	0	0	0	0	0	0	0	0	0
41	85	67	43	25	13	6	0	0	0	0	0	0	1	0	0	0	0	0	0
42	108	59	48	26	17	3	3	3	0	0	0	0	0	0	0	0	0	0	0
43	97	61	30	29	15	13	4	0	1	0	1	2	0	0	0	1	0	0	0
44	76	51	40	27	23	15	6	3	1	0	0	1	0	0	0	0	0	0	0
45	63	41	28	35	18	19	9	9	1	0	1	0	1	0	0	0	0	0	0
46	44	41	30	22	29	7	16	9	2	1	0	0	0	0	0	0	0	0	0
47	34	18	18	12	14	16	17	9	1	2	0	0	0	1	1	0	0	1	0
48	13	11	9	6	7	11	7	7	3	0	2	0	0	0	0	0	0	0	0
49	17	12	8	5	11	10	5	5	7	4	1	0	0	0	0	0	1	0	0
50	15	11	11	7	5	3	6	5	7	3	0	0	1	2	0	0	2	1	0
51	10	7	8	4	4	3	1	2	2	1	1	2	0	0	0	1	0	0	0
52	13	5	6	4	4	1	1	0	0	1	1	1	1	0	0	0	1	0	1
53	4	6	2	0	2	1	2	0	1	0	0	0	0	0	0	0	0	1	0
54	4	0	1	4	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0
55	1	0	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
56	2	0	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0
57	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

No part of one cluster becomes a part of another cluster when the iteration ends, i.e., $T_1 \subseteq P_{j_1}$, $T_2 \subseteq P_{j_2}$, $j_1 \neq j_2$ and $T_1 \cup T_2 \subseteq Q_{j_3}$ is not possible $\forall j_1$, j_2 and j_3. At most some clusters among $P_1, P_2, \ldots, P_k$ may be split and merged among themselves as the iteration progresses. Thus $k \leqslant l$. □

Proof of Theorem 2. Observe that in the $(J+1)$th iteration no element of P_r can change its membership from P_r to P_j for $j \neq r$, $r = 1, 2, \ldots, k$. At most P_r can be subdivided for $r = 1, 2, \ldots, k$.

Let P_r be split into m_r subclusters, $m_r \geqslant 1$. If $m_r = 1$ then there is no split in P_r. That means, for P_r, let $Q_{r_1}, Q_{r_2}, \ldots, Q_{r_{m_r}}$ be such that

$$\bigcup_{j=1}^{m_r} Q_{r_j} = P_r, \quad r = 1, 2, \ldots, k,$$

$$m_r \geqslant 1 \quad \text{and} \quad \sum_{r=1}^{k} m_r = l.$$

But we know that $\sum_{r=1}^{k} m_r = k$. If $\exists \alpha$, $\alpha \in \{1, 2, \ldots, k\}$ such that $m_\alpha > 1$ then $\sum_{r=1}^{k} m_r > k$, which is a contradiction. So

$$m_r \leqslant 1 \quad \forall r = 1, 2, \ldots, k.$$

But $m_r \geqslant 1$ $\forall r = 1, 2, \ldots, k$ (as stated earlier). Therefore $m_r = 1$, $r = 1, 2, \ldots, k$. Hence there is no split in P_r's. Thus

$$\{P_1, P_2, \ldots, P_k\} = \{Q_1, Q_2, \ldots, Q_k\},$$

i.e., P_r is equal to one of $Q_1, Q_2, \ldots, Q_k$ $\forall r = 1, 2, \ldots, k$. So the clusters at the end of the Jth iteration and the $(J+1)$th iteration are the same. □

References

[1] Dubes, R. and A.K. Jain (1980). Clustering methodology in exploratory data analysis. In: M.C. Yovits, Ed., *Advance in Computers*. Academic Press, New York, 113-228.

[2] Jain, A.K. (1986). Clustering analysis. In: T.Y. Young and K.S. Fu, Eds., *Handbook of Pattern Recognition and Image Processing*. Academic Press, New York.

[3] Liu, S.T. and W.H. Tsai (1989). Moment-preserving clustering. *Pattern Recognition* 22 (4), 433–447.

[4] Wishart, D. (1969). An algorithm for hierarchical classifications. *Biometrics* 25, 165–170.

[5] Dubes, R. and A.K. Jain (1979). Validity studies in clustering methodologies. *Pattern Recognition* 11, 235–254.

[6] Forgy, E.W. (1965). Cluster analysis of multivariate data. *Abstracts in Biometrics* 21 (3), 786. Efficiency Meetings, Riverside, CA.

[7] Ling, R.F. (1973). A probability theory of cluster analysis. *J. Amer. Statist. Assoc.* 68 (3), 159-164.

[8] Matusita, K. (1956). Decision rule, based on the distance, for the classification problem. *Ann. Inst. Statist. Math.* 8, 67–77.

[9] Hathaway, R.J. and J.W. Davenport (1989). Relational duals of the *c*-means clustering algorithms. *Pattern Recognition* 22 (2), 205–212.

[10] Swain, P.H. and S.M. Davis (1978). *Remote Sensing: The Quantitative Approach*. McGraw-Hill, New York.

[11] Richards, J.A. (1986). *Remote Sensing Digital Image Analysis, An Introduction*, Springer, Berlin.

[12] *IRS Data Users Handbook*, Document no. IRS/NRSA/NDC/HB-01/86, NRSA, Hyderabad, India, Sept. 1986.

[13] Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

[14] Anderbug, M.R. (1971). *Cluster Analysis for Application*. Academic Press, New York.

[15] Murthy, C.A. and D. Dutta Majumder (1990). Consistent estimation of classes for cluster analysis. Presented at *Int. Conf. Pattern Recognition*.