非階層的クラスター分析の分割最適化法における初期配置の影響*

辻本英夫[1)]　大島将人[2)]

I 問題

近年，探索的なデータ解析法の1つとして，クラスター分析が多くの分野でもちいられている。クラスター分析は，対象間の類似性（または非類似性）に基づいて，より客観的に対象**をいくつかのクラスターに分けるというある種の分類手法の総称であり，その中には数多くの多様な手法が含まれる。この多様性がクラスター分析と呼ばれる手法群の1つの特色であるが，しかしながら現状では，計算プログラムの利用できる手法をもちいるといった使い方がされがちである（大隅，1979）。その理由の1つとして，各手法のもつ特性や適用に際しての実際的な有効性があまり明確ではないからではないかということが考えられる。したがって，多様な手法群の中からより適切な手法を選択できるように，各手法の特性や有効性を検討する必要がある。

クラスター分析は，一般に，階層的クラスター分析と非階層的クラスター分析とに大別される。このうち階層的クラスター分析が，広く利用されその特性や有効性の検討が比較的よく行なわれている（たとえば柳澤・大隅，1979）に比べると，非階層的クラスター分析についての検討は少ないと思われる。そこでここでは，非階層的クラスター分析をとりあげる。しかし，その中に含まれる数多くの手法をすべて検討することはできないので，その代表的な手法群である分割最適化法（大隅，1979）のうちのさらにいくつかの手法について，それらの手法を適用する際に問題となる点について検討し，それらの手法の実際的な有効性を判断するための資料を提供したい。

分割最適化法（以下，分割型とする）と呼ばれるクラスタリング手法にも多くの手法が提案されているが，それらの手法に共通な基本的なアルゴリズムはいたって簡単である。まず，クラスタリングを始めるにあたって初期配置を与える。初期配置の与え方は大別して2通りあり，クラスターを仮に代表する点（初期代表点）を与えるか，あるいは全サンプルをいくつかの群に分ける（初期分割する）かである。初期分割した場合には，各群の重心を計算してそれらを初期代表点とする。以下，最も近い代表点へのサンプルの割り当てによるクラスターの構成・再構成と代表点の更新とを反復する。そして，あらかじめ設定されたクラスター最適化基準（クラスターのまとまりをはかる何らかの基準）に達した時に反復計算を終了し，最終的なクラスターを得る。

このようなアルゴリズムを基本とする分割型クラスタリング手法には，共通する問題点として以下の6点をあげることができ，これらの点をどのように処理するかによって様々な手法が考えられる（大隅，1979）。

1. 初期代表点・初期分割の与え方
2. サンプルをクラスターに割り当てる際の方法と，代表点の更新の時期
3. クラスター・サイズが不均衡であるときの手当ての方法
4. 異常値に対する手当ての有無
5. クラスター数の決定法（固定か，可変か）
6. 最適化の基準とそれを達成するアルゴリズム

このうち，クラスター数が固定か可変かによって，分割型クラスタリング手法は2分される。クラスター数固定

---

* The influence of initial configurations on clusters obtained by partitioning-optimization techniques of nonhierarchical clustering methods.

本研究のデータ分析のための計算は，名古屋大学大型計算機センターのFACOM M－200によった。

1）名古屋大学大学院教育学研究科博士課程（後期課程）教育心理学専攻

2）岐阜県立大垣商業高等学校

** クラスタリングの対象については，サンプルを分類する，変数を分類する，サンプルと変数を同時に分類するという3つの場合があるが，後述する林（1978），久世他（1979）及びここでの分析においては，いずれもサンプルを分類するので，クラスタリングの対象をサンプルに限定して論をすすめる。

の場合にはクラスター数を，また，可変の場合にはサンプンのクラスターへの合併・分離あるいはクラスター同士の合併・分離のための基準を，それぞれ最初に指定しなければならない。クラスター分析は探索型の研究のための手法としてもちいられることが多いので，クラスタリングに際しての情報が少ないという状況を想定するならば，クラスター数を指定する方がより困難が少ないと思われる。また，クラスター数をかえて分析を反復することもより容易であろう。したがって本研究では，こういった点でより実用的であると思われるクラスター数固定の分割型クラスタリング手法に問題を限定し，そのような手法として，Anderberg（1973）に紹介されているForgy法，Jancey法，MacQueen's K-means法及びconvergent K-means法の4つの手法をとりあげる。そして，これらの手法をもちいる際の主要な問題の1つである初期代表点・初期分割の与え方（以下，初期配置法とする）をとりあげて，検討を加える。

すなわち，本研究の主目的は，クラスター数固定の分割型クラスタリング手法について，初期配置に関する諸問題を検討することである。具体的には，以下の4点を問題とする。

1. 初期配置の違いが，結果として得られるクラスターに影響を及ぼすか。
2. 及ぼすならば，クラスタリング手法間に初期配置の影響に対する敏感性に違いがあるか。つまり，クラスタリング手法間に初期配置の違いによるクラスターの影響のされやすさの点で違いがあるか。
3. また，初期配置法に明らかに優れたものがあるか。
4. あるとすれば，そのような初期配置法はクラスタリング手法によって異なるか。

さらにこの4点に加えて，次の2点についても検討する。

5. データにおけるサンプルの順序は，MacQueen's K-means法やconvergent K-means法をもちいた場合に，結果として得られるクラスターに影響を及ぼすか。（Forgy法・Jancey法では，すべてのサンプルが割り当てられるまで代表点は固定されているので，サンプル順序の影響はない。）
6. 4つのクラスタリング手法の中で，明らかに他の手法より優れているものがあるか。

Ⅱ 方　法

1. データ

HAYASHI－1，HAYASHI－2，KUZEと名づけた3種類のデータをもちいる。その内容については，**表1**に示した通りである。ただし，データKUZEでのクラスター数3は，主要なクラスターが3個ということであり，どのクラスターにも属さないサンプルが全サンプル94名中29名ある。

2. 距離測度

各サンプルとクラスターの代表点（重心）との距離として，ユークリッド距離の2乗をもちいる。

3. クラスターの概念

クラスターをどういうものと考えるかについては様々な考え方があるが，ここでは多次元の球あるいはそれに近い集塊状をなすサンプルの集りをクラスターと考える。

4. クラスター数

林（1978），久世他（1979）でそれぞれ得られたクラスター数*（**表1**参照）を，各データをクラスタリングする際のクラスター数としてそのままもちいる。

5. クラスタリング手法

もちいたクラスタリング手法は，Forgy法・Jancey法・MacQueen's K-means法・convergent K-means法の4つで，1）クラスター・サイズが不均衡な場合でも何ら手当てを行なわない，2）異常値に対する手当て

表1　もちいたデータ

データ名	HAYASHI－1	HAYASHI－2	KUZE
内容	林（1978）の相貌尺度についての平均評定値	林（1978）の性格尺度についての平均評定値	久世他（1979）の6年間の3つの社会的態度得点
サンプル数	刺激図形（12人の平均評定値）×90		94（男子50，女子44）
変数数	20		18
行なわれているクラスタリング	Ward法		centroid法
クラスター数	6	7	3

* 林（1978），久世他（1979）ではともに，各クラスターに含まれるサンプル数，各クラスターの変数ごとの平均値，樹状図を参考にしてクラスター数を決定している。

も行なわない，3）クラスター数を固定してクラスタリングを行なう，4）クラスター内平方和の最小化を最適化基準とするという4点で共通で，いずれも非階層的分割型のクラスタリング手法である。加えてこれらの4つの手法は，代表点の更新の方法及び更新の時期によって，それぞれ2つのグループに分けることができる。更新の方法では，Jancey 法以外の3つの手法が共通した方法をとっている。また，更新の時期については，Forgy 法と Jancey 法，MacQueen's K-means 法と convergent K-means 法とでそれぞれ共通している。

Forgy 法：Forgy（1965）によって提案された方法は，以下のステップからなる単純なアルゴリズムによってクラスタリングを行なう方法である。

ステップ1；クラスタリングを始めるにあたって，初期代表点を与えるか初期分割する。初期分割を与えた場合には，各群の重心を算出して初期代表点とする。

ステップ2；サンプルを順次最も近い代表点に割り当てることによってクラスタリングを行なう。すべてのサンプルがいずれかの代表点に割り当てられるまで代表点は固定される。

ステップ3；各代表点に割り当てられたサンプルに基づいて各クラスターの重心を再計算し，新しい代表点とする。

ステップ4；ステップ2においてクラスターの成員が変化しなくなるまで，ステップ2と3をくり返す。

Jancey 法：Jancey（1966）の方法は，ステップ3以外はまったく Forgy 法と同じである。ステップ3（代表点の更新）は以下のように行なわれる（**図1** 参照）。

ステップ3；ステップ2で各代表点に割り当てられたサンプルに基づいてクラスター重心を再計算し，この再計算された重心（C_{kj}）と再計算される前の代表点（S'_{kj}）とをもちいて，次式により新しい代表点（S_{kj}）を算出する。

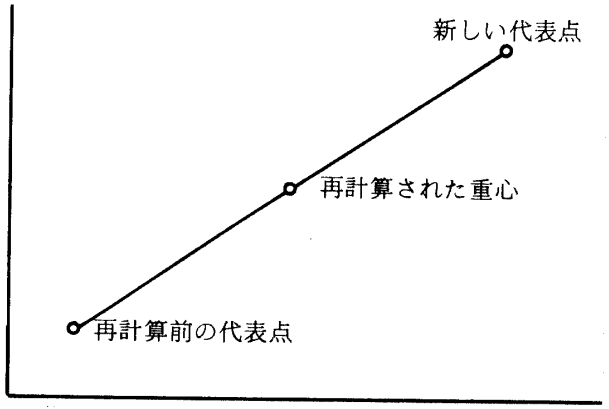

図1　Jancey 法での代表点の更新
（Anderberg，1973）

$$S_{kj}=2C_{kj}-S'_{kj}$$

ここで k, j はそれぞれクラスターと変数を示す。

MacQueen's K-means 法：MacQueen（1967）の提案した方法は，Forgy 法や Jancey 法と異なり，サンプルがクラスターに割り当てられるごとに代表点を更新するという点に特徴がある。MacQueen の方法では，以下のステップによってm個のサンプルがk個のクラスターに分割される。

ステップ1；データの最初のk個のサンプルをそれぞれクラスターの代表点とする。

ステップ2；残りの（m－k）個のサンプルを順次最も近い代表点に割り当てる。この時，サンプルが割り当てられるごとに重心を再計算し，代表点を更新する。

ステップ3；ステップ2においてすべてのサンプルが代表点に割り当てられたならば，代表点を固定したままで，再度すべてのサンプルを順次最も近い代表点に割り当てて最終的なクラスターを得る。

convergent K-means 法：convergent K-means 法は MacQueen's K-means 法の変法であり，以下のステップによってクラスタリングが行なわれる。

ステップ1；クラスタリングを始めるにあたって，初期代表点を与えるか初期分割を与える。初期代表点を与えた場合には，サンプルを順次最も近い代表点に割り当て，各群の重心を計算して代表点とする。初期分割を与えた場合にも，各群の重心を算出して代表点とする。

ステップ2；サンプルを順次最も近い代表点に割り当てる。この時，割り当てられたクラスターが割り当て前にそのサンプルが属していたクラスターと異なるならば，割り当て前と割り当て後に属するクラスターの重心を再計算して，各々の代表点を更新する。

ステップ3；クラスター成員が変化しなくなるまで，ステップ2をくり返す。

6. 初期配置法

分割型クラスタリング手法の多くは初期配置の与え方を指定しているが，一般に利用者の便宜のためにそうされているのであり，指定された方法をもちいなければならないという理由はない。

ここでは，**表2** に示す7つの初期配置法をとりあげる。ここでとりあげた方法が初期配置の与え方のすべてを尽くしているというわけではないが，機械的に選ぶ方法としてＩＣ1を，ランダムに選ぶ方法としてＩＣ2・3を，何らかの他の情報に基づいて選ぶ方法としてＩＣ4・5・6・7をというように，各種の代表的な方法を含むよ

う配慮した。

表2 もちいた初期配置法

IC1	データの最初のk†サンプルを初期代表点とする
IC2	ランダムに選んだkサンプルを初期代表点とする
IC3	k個の初期代表点の変数値として乱数をもちいる
IC4	非線型マッピングの結果から主観的にkサンプルを選んで初期代表点とする
IC5	階層的クラスター分析の結果に基づいて初期分割する
IC6	主要な変数の分布に基づいて初期代表点を選択する
IC7	主要な変数の分布に基づいて初期代表点を選択する

†kはクラスター数を表わす。

なお，IC5について，階層的クラスター分析の結果としては，林（1978），久世他（1979）の分析結果をそのままもちいる。同じくIC5について，MacQueen's K-means法の場合には，プログラムの制約*上，どのデータをもちいる場合でも，初期分割のかわりに，各クラスターの成員の中からそれぞれ1サンプルをランダムに選んでそのクラスターの代表点とし，データKUZEをもちいる場合には，データの制約**上，どのクラスタリング手法の場合にも，同様に初期代表点を選んで分析を行なう。またIC3については，MacQueen's K-means法の場合には，プログラムの制約*上除外する。

IC6とIC7については，具体的には次のように初期代表点を与える。まずHAYASHI－1と－2については，林（1978）が因子分析も行なっているので，その結果得られた因子にそれぞれ高い負荷を示す変数2・8・9・20（HAYASHI－1）と変数5・7・8（HAYASHI－2）を主要変数として選ぶ。KUZEについては，6年目の3つの態度得点である変数6・12・18を主要変数とする。次に，単峰形分布ならば1つ，双峰形分布ならば2つというように，各主要変数の分布に基づいてサンプルをいくつかのグループに分ける。そして，それらのグループの組み合わせでさらにサブ・グループに分け，得られたサブ・グループの中から，グループ・サイズの大きい順（IC6）ないしは小さい順（IC7）にクラスター数だけサブ・グループを選択し，選択した各サブ・グループからランダムに1サンプルずつを選んで初期代表点とする。

* MacQueen's K-means法のプログラムは，クラスター数分だけデータの最初からサンプルを選んで，それらを初期代表点とするようになっている。

** データKUZEは久世他（1979）の全サンプルを含むので，久世他（1979）で得られたクラスターのどれにも属さないサンプルをも含む。

7. 測　度

クラスター成員の一致度を示す測度としてCramérのV係数を，クラスターのまとまりの程度を示す測度としてクラスター内平方和をもちいる。CramérのV係数は0～1の値をとり，値が大きいほど一致度が高いことを示す。また，クラスター内平方和については，その値が小さいほどクラスターのまとまりが良いことを示す。それぞれの計算式は以下の通りである。

1) CramérのV係数

$$V=\left(\frac{\phi}{min\{(r-1),(c-1)\}}\right)^{\frac{1}{2}}$$

ここで，$\phi=(\chi^2/N)^{\frac{1}{2}}$，$N$はサンプル総数，$r$は一方のクラスター数，$c$は他方のクラスター数

2) クラスター内平方和

$$\sum_{k=1}^{m}\left(\frac{1}{n_k}\sum_{\substack{\alpha,\beta\in C_k\\ \alpha<\beta}}^{n} d_{\alpha\beta}^2\right)$$

ここで，αとβはクラスターC_kの成員，dはαとβの距離，n_kはクラスターC_kの成員数，mはクラスター数

8. プログラム

クラスタリングのプログラムは，Anderberg（1973, Pp. 306－325）のプログラムを利用しやすいよう書き直してもちいる。クラスター内平方和を算出するサブルーチン，及びクラスタリングの途中でクラスター・サイズが0となるクラスターが生じた場合に対する手当てのためのサブルーチンを追加した以外は，内容的には変更はない。クラスター・サイズが0となるクラスターが生じた場合の手当てとしては，以下の処置を行なう。

1) サイズ0のクラスターが1つの場合；
   他のクラスターの更新後の代表点の重心をサイズ0のクラスターの代表点とする。
2) サイズ0のクラスターが2つ以上の場合；
   1）の仕方でまず1つのクラスターの代表点を決定し，そのクラスターも含めて同様に2番目のクラスターの代表点を決定する。それ以降も同様に順次決定していく。
3) サイズ0のクラスターの数がクラスター総数から1引いた数以上の場合には，上記の仕方で決定できないので，計算を打ち切る。

Ⅲ　結果と考察

最初に，クラスタリングの途中でクラスター・サイズ

が0となったものについて記しておく。データＨＡＹＡＳＨＩ－１とＫＵＺＥについてはJancey法をもちいた際に，またＨＡＹＡＳＨＩ－２についてはMacQueen's K-means法以外の手法をもちいた際に，いずれもＩＣ３によって初期代表点を与えた場合にクラスター・サイズが0となるクラスターが計算途中で生じた。どの場合にも，「プログラム」の項で述べた処置を行なって，最後まで計算を実行した。

表3は，初期配置の影響をみるために，２つの初期配置法によって得られるクラスター間のCramérのＶ係数を，データ及びクラスタリング手法ごとにまとめたものである。データＫＵＺＥで，Jancey法をもちいた場合のＩＣ２とＩＣ５によって得られるクラスター間でのみ全く同一の成員をもつクラスターが得られる以外は，程度の違いはあるが，結果として得られるクラスターは初期配置の影響を受ける。ここでもちいたデータ・クラスタリング手法・初期配置法については，データＫＵＺＥにMacQueen's K-means法をもちいた場合のＩＣ１とＩＣ６によって得られるクラスター間で最もクラスターの成員の一致度が低く，そのＶ係数は.475である。この値がどの程度の違いを表わしているのかを示すために，この場合のクロス表を**表4**に示す。明らかにＩＣ１をもちいた場合とＩＣ６をもちいた場合とでは，得られるクラスターが全く異なっている。

表３－１－１　Cramérのｖ係数
（データ：HAYASHI－１）
（手　法：Forgy）

ＩＣ	1	2	3	4	5	6
2	.834					
3	.733	.697				
4	.655	.732	.676			
5	.802	.858	.663	.772		
6	.721	.841	.687	.764	.799	
7	.771	.873	.691	.706	.840	.831

表３－１－２　Cramérのｖ係数
（データ：HAYASHI－１）
（手　法：Jancey）

ＩＣ	1	2	3	4	5	6
2	.714					
3	.735	.874				
4	.835	.837	.839			
5	.812	.834	.818	.974		
6	.738	.839	.897	.872	.851	
7	.814	.859	.872	.953	.927	.818

表３－１－３　Cramérのｖ係数
（データ：HAYASHI－１）
（手　法：MacQueen's K-means）

ＩＣ	1	2	3	4	5	6
2	.590					
3	……	……				
4	.561	.730	……			
5	.557	.912	……	.693		
6	.559	.840	……	.819	.812	
7	.668	.717	……	.746	.701	.693

表３－１－４　Cramérのｖ係数
（データ：HAYASHI－１）
（手　法：convergent K-means）

ＩＣ	1	2	3	4	5	6
2	.802					
3	.731	.715				
4	.654	.695	.722			
5	.802	.828	.678	.727		
6	.721	.827	.724	.758	.799	
7	.778	.893	.747	.708	.827	.868

表３－２－１　Cramérのｖ係数
（データ：HAYASHI－２）
（手　法：Forgy）

ＩＣ	1	2	3	4	5	6
2	.847					
3	.822	.880				
4	.828	.804	.824			
5	.821	.863	.917	.846		
6	.828	.765	.821	.882	.840	
7	.742	.773	.807	.754	.800	.725

表３－２－２　Cramérのｖ係数
（データ：HAYASHI－２）
（手　法：Jancey）

ＩＣ	1	2	3	4	5	6
2	.846					
3	.730	.708				
4	.931	.853	.741			
5	.852	.828	.776	.851		
6	.842	.836	.776	.840	.988	
7	.785	.699	.795	.778	.789	.786

表3－2－3 Cramér のV係数
（データ：HAYASHI－2）
（手　法：MacQueen's K-means）

IC	1	2	3	4	5	6
2	.735					
3	……	……				
4	.758	.746	……			
5	.790	.783	……	.861		
6	.848	.824	……	.762	.796	
7	.688	.577	……	.682	.725	.648

表3－2－4 Cramér のV係数
（データ：HAYASHI－2）
（手　法：convergent K-means）

IC	1	2	3	4	5	6
2	.857					
3	.816	.764				
4	.870	.811	.894			
5	.830	.863	.835	.834		
6	.837	.768	.845	.879	.839	
7	.733	.801	.758	.812	.736	.720

表3－3－1 Cramér のV係数
（データ：KUZE）
（手　法：Forgy）

IC	1	2	3	4	5	6
2	.847					
3	.459	.503				
4	.985	.836	.466			
5	.851	.941	.516	.841		
6	.647	.598	.474	.647	.608	
7	.862	.983	.504	.851	.923	.622

表3－3－2 Cramér のV係数
（データ：KUZE）
（手　法：Jancey）

IC	1	2	3	4	5	6
2	.795					
3	.655	.588				
4	.930	.868	.610			
5	.795	1.000	.588	.868		
6	.555	.595	.455	.527	.595	
7	.970	.797	.673	.926	.797	.546

表3－3－3 Cramér のV係数
（データ：KUZE）
（手　法：MacQueen's K-means）

IC	1	2	3	4	5	6
2	.615					
3	……	……				
4	.576	.580	……			
5	.609	.924	……	.552		
6	.475	.622	……	.616	.589	
7	.481	.697	……	.504	.698	.582

表3－3－4 Cramér のV係数
（データ：KUZE）
（手　法：convergent K-means）

IC	1	2	3	4	5	6
2	.821					
3	.483	.546				
4	.985	.836	.481			
5	.800	.974	.570	.815		
6	.691	.825	.635	.701	.847	
7	.851	.966	.545	.867	.938	.797

表4 IC1と6によって得られた各クラスターの成員†

IC	6				
	クラスター	1	2	3	計
1	1	26	0	0	26
	2	18	4	27	49
	3	19	0	0	19
	計	63	4	27	94

† データ：KUZE
手　法：MacQueen's K-means

この初期配置の影響をさらにクラスターのまとまりの良さという点から調べるために，**表5**にクラスター内平方和を示す。クラスター内平方和でみると，初期配置法の違いによって最もクラスターに違いがあるのは，データHAYASHI－1にMacQueen's K-means法をもちいた場合のIC1とIC2によって得られるクラスターの間であり，クラスター内平方和で195.381の差がある。この差は，全体平方和*（データHAYASHI－1では，1838.561）の10.6％にあたり，無視できる差であるとは言い難い。

* 全体平方和は次式で計算される；

$$\sum_{j=1}^{m}\left(\sum_{i=1}^{n}x_{ij}^{2}-n\bar{x}_{j}^{2}\right)$$

ここで $i=1,2,\cdots,n$ ；サンプル
$j=1,2,\cdots,m$ ；変数

各データについての全体平方和は以下の通り。

HAYASHI－1	1838.561
HAYASHI－2	3102.557
KUZE	45861.937

表5－1　クラスター内平方和（データ：HAYASHI－1）

手法	Forgy	Jancey	MacQueen's K-means	convergent K-means	平均	SD
IC						
1	712.466	713.593	869.940	712.466	752.116	68.027
初 2	658.224	677.073	674.559	661.808	667.916	8.050
期 3	747.466	678.243	……	744.207	723.305	31.892
配 4	732.841	656.626	728.625	737.883	713.994	33.283
置 5	670.386	658.801	681.172	670.386	670.186	7.912
法 6	685.624	664.536	690.250	685.624	681.509	9.979
7	671.758	660.120	742.227	667.755	685.465	33.037
平均	696.966	672.713	731.129	697.161	698.321	
SD	31.733	18.511	66.789	31.781		44.868

表5－2　クラスター内平方和（データ：HAYASHI－2）

手法	Forgy	Jancey	MacQueen's K-means	convergent K-means	平均	SD
IC						
1	901.005	878.228	912.307	898.925	897.616	12.298
初 2	904.773	893.082	1026.134	904.773	932.191	54.448
期 3	894.925	1133.758	……	893.400	974.028	112.948
配 4	879.159	885.463	914.692	891.112	892.607	13.434
置 5	883.245	885.284	883.377	883.245	883.788	.866
法 6	937.283	884.289	897.338	938.940	914.463	24.102
7	1004.096	954.977	1082.479	974.970	1004.131	48.489
平均	914.927	930.726	952.721	912.195	926.713	
SD	40.410	86.399	74.366	30.526		63.831

表5－3　クラスター内平方和（データ：KUZE）

手法	Forgy	Jancey	MacQueen's K-means	convergent K-means	平均	SD
IC						
1	30049.602	30123.691	31885.340	30044.219	30525.713	785.609
初 2	29896.844	29902.223	29962.551	29902.223	29915.960	26.986
期 3	32006.969	33542.066	……	32359.570	32636.202	656.518
配 4	30044.219	30024.977	32720.066	30020.875	30702.534	1164.856
置 5	30045.117	29902.223	29926.828	29934.055	29952.056	55.011
法 6	32113.352	32398.453	32537.426	30250.895	31825.032	921.595
7	29894.367	30104.598	32650.258	29894.367	30635.898	1166.154
平均	30578.639	30856.890	31613.745	30343.743	30819.903	
SD	939.504	1373.471	1210.880	830.736		1198.719

次に，クラスタリング手法間に初期配置の影響に対する敏感性の点で違いがあるかどうかをみるために，手法ごとにV係数の平均をとったものが**表6**である。**表6**から，どのデータについても一貫して，MacQueen's K-means法が他の3手法に比べて初期配置の影響に敏感であることがわかる。他の手法については，V係数の高い順に並べると，データHAYASHI－1ではJancey法－convergent K-means法－Forgy法という順であり，

表6　クラスタリング手法ごとのCramérのV係数の平均・SD

データ / クラスタリング手法	HAYASHI－1		HAYASHI－2		KUZE	
	平均	SD	平均	SD	平均	SD
Forgy	.759	.068	.819	.047	.713	.183
Jancey	.843	.064	.811	.067	.721	.161
MacQueen's K-means	.707	.105	.748	.074	.608	.105
convergent K-means	.762	.062	.814	.049	.761	.158

データHAYASHI－2，KUZEではそれぞれ，Forgy法－convergent K-means法－Jancey法，convergent K-means法－Jancey法－Forgy法と一定しておらず，また，手法間のV係数の平均の差も小さい（最大で，データHAYASHI－1のJancey法・convergent K-means法間の .081）ので，初期配置の影響に対する敏感性の点ではあまり差はないと思われる。

初期配置法の優劣についての結果は，**図2**・**図3**に示す。**図2**は，**表5**に行平均として示した初期配置法ごとのクラスター内平方和の平均の小さい順に各初期配置法を並べたものであり，初期配置法間の数値はクラスター内平方和の平均の差の全体平方和に対する割合（%）を示している。**図3**は同様に，**表5**に示したクラスター内

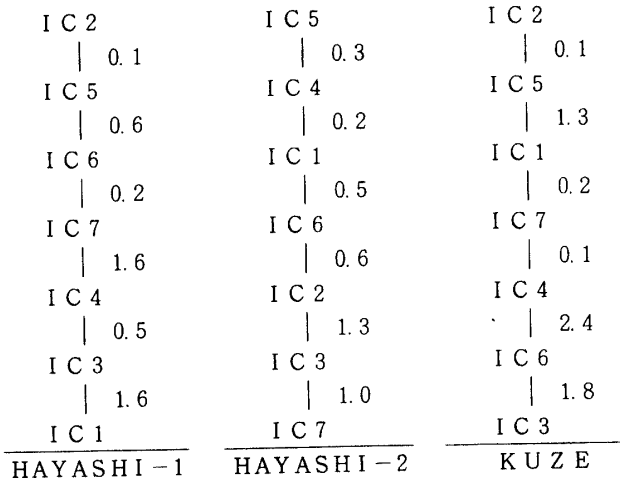

図2　クラスター内平方和の平均値による初期配置法の順位

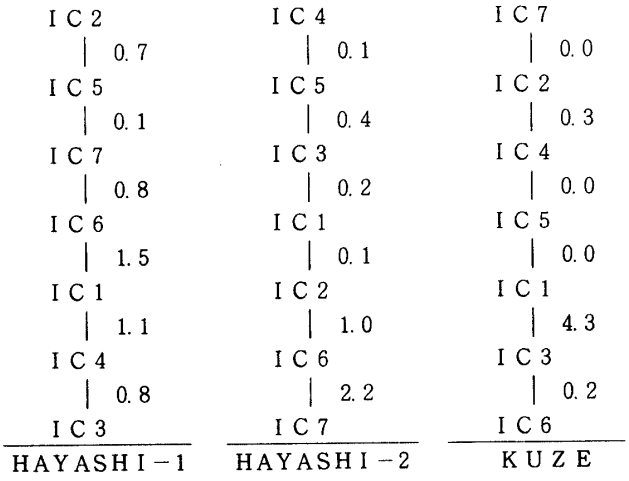

図3－1　Forgyでのクラスター内平方和による初期配置法の順位

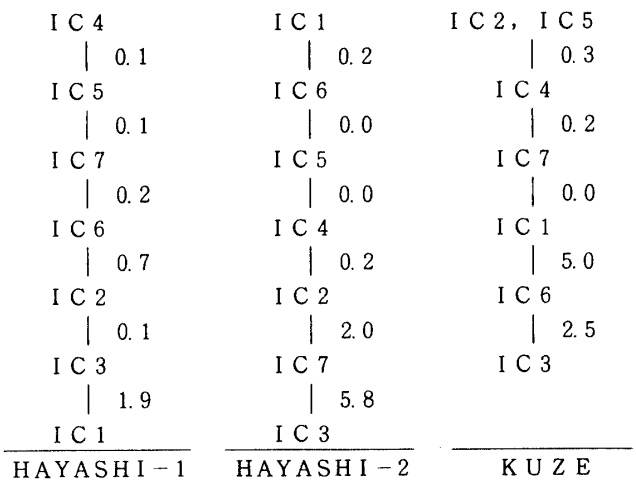

図3－2　Jancey法でのクラスター内平方和による初期配置法の順位

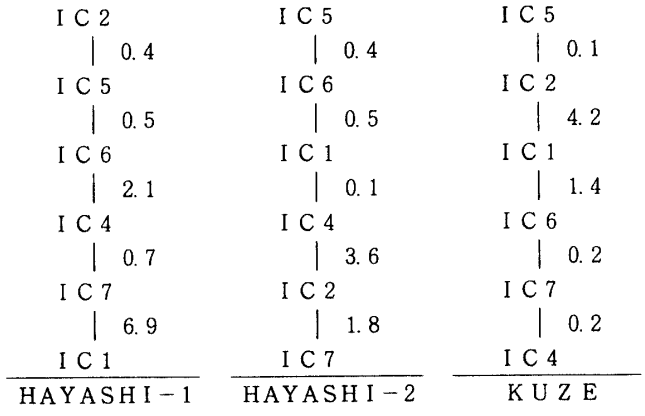

図3－3　MacQueen's K-means法でのクラスター内平方和による初期配置法の順位

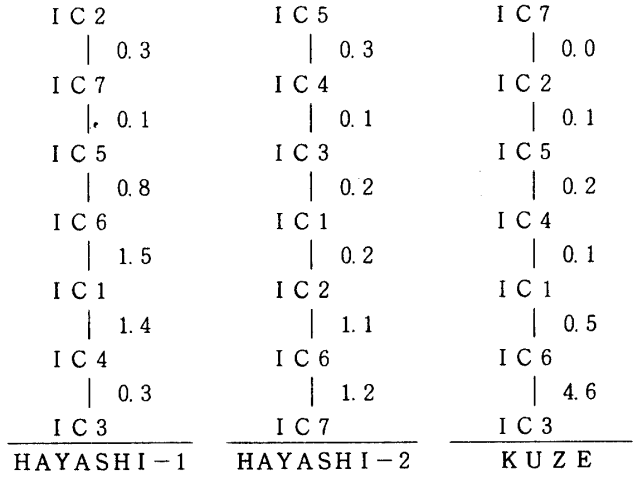

図3－4　convergent K-means法でのクラスター内平方和による初期配置法の順位

平方和の小さい順に，クラスタリング手法ごとに初期配置法を並べたものである。

図2からは，まず，どのデータについても一貫して，ＩＣ５が他のほとんどの初期配置法よりも良くまとまったクラスターを与えることがわかる。データＨＡＹＡＳＨＩ－２の場合には最良のまとまりをもつクラスターを与えるし，ＨＡＹＡＳＨＩ－１，ＫＵＺＥの場合にも２番目に良いまとまりをもつクラスターを与え，しかも，その場合に最良のまとまりをもつクラスターを与えるＩＣ２とのクラスター内平方和の平均の差も全体平方和の0.1%にすぎない。**図2**からはまた，ＩＣ５とは対照的にＩＣ３をもちいた場合には，どのデータについても，他の方法よりもまとまりの悪いクラスターが生じるという傾向がうかがえる。すなわち，ＩＣ３による場合には，不適切な初期代表点が選ばれやすいと思われる。この点については，計算途中でクラスター・サイズが０となるクラスターが生じたのが，すべてＩＣ３の場合であるということからも認められる。他の初期配置法については，データによって結果が異なり一定の傾向がうかがえないので，その優劣について判断できない。

このような傾向は，クラスタリング手法ごとに初期配置法の優劣をみても同様に認められる（**図3**）。**図3**からも，順位の変動は多少大きくはなるが，どの手法どのデータについてもほぼ一貫してＩＣ５が良い。また，データＨＡＹＡＳＨＩ−２にForgy法とconvergent K-means法をもちいた場合に比較的良い結果をもたらすが（クラスター・サイズ０のクラスターが生じた場合のプログラム上の処理が，たまたま良い結果をもたらしたのではないかと思われる），ＩＣ３は全体的に，他の初期配置法に比べてあまり良いまとまりをもったクラスターを与えない。特にJancey法では，この傾向が顕著である。特定のクラスタリング手法にのみ優れた初期配置法というものは，**図3**からはほとんど認められない。ただ，１つの目安として，全体平方和に対する割合１%までの差を差として問題にしないとするならば，Forgy法・convergent K-means法については，ＩＣ２も比較的良い方法といえるかもしれない。

副次的な目的であるMacQueen's K-means法・convergent K-means法におけるサンプル順序の影響については，両手法とも，アルゴリズムのステップ２でサンプル順序をまったく逆にするという形で再分析を行なった。データはＨＡＹＡＳＨＩ－１のみをもちいた。この再分析の結果得られたクラスター内平方和を，サンプル順序が正順の場合の結果と並べて**表7**に示す。その結果からは，サンプル順序が正順と逆順の場合ではクラスター内平方和に違いがみられ，どちらの手法についても，データにおけるサンプル順序もまた，結果として得られるクラスターに影響を及ぼす要因であることがわかる。しかしながら差はそれほど大きくなく，初期配置の違いよりもその影響は小さいと思われる。さらにconvergent K-means法については，ＩＣ５とＩＣ６で同一の値を示していることが注目される。そこでこれらの場合について，データＨＡＹＡＳＨＩ－２についても同じく再分析を行なってみると，ＩＣ５については同様に，正順の場合と逆順の場合でのクラスター内平方和が一致した。このことは，convergent K-means法でＩＣ５によって初期配置を与える場合には，データにおけるサンプル順序の影響がほとんどないことを予想させる。また，MacQueen's K-means法についても，ＩＣ５の場合の正順・逆順間のクラスター内平方和の差は０ではないけれども（これは初期分割でなく初期代表点を与えたためでもあると考えられる），他の初期配置法に比べてかなり小さい。したがってこれらの結果から，サンプル順序の影響という点についても，他の初期配置法に比べてＩＣ５が優れていると言えるように思われる。

表7　正順・逆順の場合のクラスター内平方和
（データ：HAYASHI−1）

初期配置法＼手法・順序	MacQueen's K-means 正順	MacQueen's K-means 逆順	convergent K-means 正順	convergent K-means 逆順
ＩＣ１	869.940	850.919	712.466	724.983
2	674.559	764.947	661.808	658.655
3	……	……	744.207	748.672
4	728.625	746.395	737.883	743.025
5	681.172	676.574	670.386	670.386
6	690.250	716.286	685.624	685.624
7	742.227	752.364	667.755	665.899
平　均	731.129	751.248	697.161	699.606
ＳＤ	66.789	53.143	31.781	35.454

最後に，ここでとりあげた４つのクラスタリング手法の優劣について述べる。**図4**は，**表5**に列平均として示したクラスタリング手法ごとのクラスター内平方和の平均について，その値の小さい順に各手法を並べたものである。手法間の数値は，クラスター内平方和の平均の差の全体平方和に対する割合（%）を示している。図には，どのデータについても一貫して，MacQueen's K-means法に他の手法より大きなクラスター内平方和が認められる。すなわち，MacQueen's K-means法で得られるクラスターは，他の手法で得られるクラスターに比べてまとまりが悪いと言えそうである。さらに，他の手法より

も初期配置の影響に対して敏感であること，得られるクラスターがサンプル順序に影響を受けることをも考え合わせると，MacQueen's K-means 法は他の3つの手法よりも劣ると言えよう。他の3つの手法については，**図7**からも，先に述べた初期配置の効果に対する敏感性という点からも，優劣の判断はつけ難い。

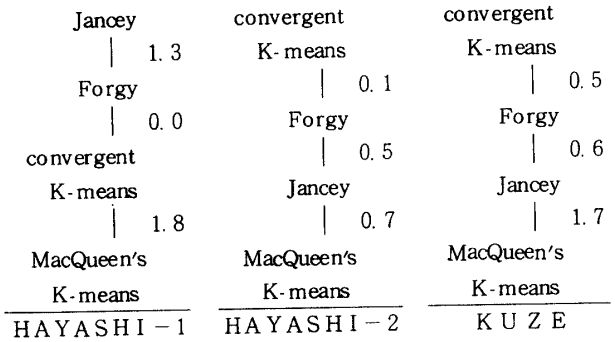

図4　クラスター内平方和の平均値による
クラスタリング手法の順位

Ⅳ　まとめ

本研究の結果は，以下のようにまとめられる。ただ，本研究ではわずかに3つのデータしかもちいなかったし，これらのデータがデータの種々のタイプを代表するとも言い難いので，ここでの結論がどの範囲まで一般化できるかという点には問題が残る。

1. 初期配置の違いは，結果として得られるクラスターに影響を及ぼす。その影響の程度は，もちいるデータ・クラスタリング手法によっても異なるが，初期配置法の選び方によっては，クラスター内平方和で全体平方和の10％以上も違ってくる可能性がある。

2. クラスタリング手法の初期配置の影響に対する敏感性については，MacQueen's K-means 法が他の3つの手法よりも敏感であることがうかがえた。

3. 初期配置法の優劣については，ここでもちいたどのデータにおいても，階層的クラスタリングの結果に基づいて初期分割するという方法（ＩＣ５）が比較的まとまりの良いクラスターを与えることが認められた。逆に，初期代表点の変数値に乱数をもちいるという方法（ＩＣ３）は，多くの場合，適切な初期配置を与えないことが示された。

4. このことは，クラスタリング手法ごとに検討しても，ここでもちいたどの手法についてもほぼ認められた。特定の手法にのみ優れた初期配置法としては，特に顕著なものはなかった。

5. MacQueen's K-means法とconvergent K-means法では，データにおけるサンプル順序も結果として得られるクラスターに影響を与えることが示された。ただ，影響の程度はそれほど大きくなく，ＩＣ５によって初期配置を与える場合には，特に影響が少ないと思われる。

6. クラスターのまとまりの良さ・初期配置の影響に対する敏感性・データにおけるサンプル順序の影響といった点から判断すると，MacQueen's K-means 法は他の3つの手法よりも劣ると結論づけられる。他の手法については，ここでの結果からは，明確に優劣をつけることはできなかった。

以上の結果から，クラスター数固定の分割型クラスタリング手法（分割最適化法）でクラスタリングする場合は，まず階層的クラスター分析を行ない，その結果から初期分割を決定して，Forgy, Jancey, convergent K-means 法のいずれかの手法でクラスタリングを行なえば良いと思われる。もちろん，時間的に余裕があるならば，3つの手法をすべて行ない，できるならば初期配置についてもいくつかの初期配置に基づいてクラスタリングを行なって，その中で最も良い結果を採用すべきであろう。

どのような階層的クラスター分析を前もって行なうか，加えて，分割型クラスタリング手法を併用してより良い結果が得られるかという点は，今後検討すべき問題の1つである。ここでもちいたデータに関して言えば，ＨＡＹＡＳＨＩ－１・ＨＡＹＡＳＨＩ－２については，林（1978）が行なっている手法はWard 法で，その結果のクラスター内平方和は681.966（ＨＡＹＡＳＨＩ－１）と914.993（ＨＡＹＡＳＨＩ－２）であり，どちらのデータについても，ここでＩＣ５によって初期配置を与えて行なった分割型クラスタリング手法での結果の方が，より小さなクラスター内平方和を示している。ＫＵＺＥに関しては，久世他（1979）で行なっている手法はcentroid 法で，その結果のクラスター内平方和は14378.617であり，一方，久世他（1979）が得た3クラスターに属しているサンプルのみについてＩＣ５をもちいてクラスタリングを新たに行なった結果では，MacQueen's K-means 法の場合に14367.066，他の3手法の場合にはすべて14093.219というように，やはりより小さなクラスター内平方和が示される。したがってこれらの場合については，クラスターのまとまりという点では，分割型クラスタリングを併用した方がより良い結果が得られたと言える。

また，先にも述べたように，ここでは3つのデータについて分析を行なったにすぎないので，さらに多くの種々のタイプのデータについて同様な分析を行ない，資料を蓄積していく必要があろう。この点に関連して，ここではとりあげられなかった問題で，「初期配置法・クラスタリング手法の違いによって，クラスターに組織的な違

いが生じるのか。もし生じるのならばどのような違いか。」という問題については，今後の課題として検討してみたい。加えて，ここでとりあげたクラスタリング手法は，クラスター数固定の分割型クラスタリング手法4種にすぎず，はじめに述べたように，この他にも数多くの手法が提案されているので，それらの手法についても，その特性や実際的な有効性がさらに検討され明らかにされることを期待したい。

付　記

本研究のために快くデータをお借し下さった久世敏雄教授をはじめとする皆さん，林文俊さん，及び御指導いただいた内田良男教授に深く感謝いたします。

なお，本稿は，昭和54年度の教育学研究科の授業に関連して，われわれが行なった検討結果をまとめたものである。

文　献

Anderberg, M.R. 1973 *Cluster analysis for applications.* Academic Press.

Forgy, E.W. 1965 Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics,* **21**, 768. (Abstract)

林　文俊　1978　相貌と性格の仮定された関連性(3)―漫画の登場人物を刺激材料として―　名古屋大学教育学部紀要―教育心理学科―，**25,** 41―45.

Jancey, R.C. 1966 Multidimensional group analysis. *Australian Journal of Botany,* **14**, 127 – 130.

久世敏雄・後藤宗理・二宮克美・宮沢秀次・池田博和・伊藤義美・浅野敬子　1979　中学生・高校生の社会的態度に関する縦断的研究(1)　名古屋大学教育学部紀要―教育心理学科―，**26,** 17―35.

MacQueen, J.B. 1967 Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* **1**, 281 – 297.

大隅　昇　1979　クラスター分析はどう使われるか　数理科学，No.190, 26―34.

柳澤幸雄・大隅　昇　1979　Single linkage 法と Complete linkage 法の特性とクラスター数評価基準　応用統計学，**8,** 51―71.

(1980年7月31日　受稿)