北海道大学工学部研究報告
第77号 (昭和50年)

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 77 (1975)

線形分離不可能なデータに関する Clustering の手法について

佐藤義治* 河口至商*

(昭和50年3月31日受理)

On Clustering Techniques for Linearly Non-separable Data Units

Yoshiharu SATO Michiaki KAWAGUCHI

(Received March 31, 1975)

Abstract

The clustering techniques referred to as the combinatorial method by G. N. Lance and W. T. Williams are regarded to be useful for practical applications. But in this method the clusters are not uniquely determinant and this method is insufficient for linearly non-separable cases. Here we offer a new clustering technique with regard to factor analysis and a similarity coefficient.

Further, in comparison, regarding the linearly non-separable data units we have discussed the characteristic features of clustering techniques, including the combinatorial method, and the new technique proposed here. And we have shown that the new clustering technique is more useful for such data units as compared with the combinatorial method.

1. 序　　論

クラスター分析といわれている手法は，データ解析における手法の1つであり，その目的は，与えられたデータをいくつかのグループに分類することによって，そのデータに内在する構造を明らかにすることである。すなわち，データを構成する個々の事象あるいは個体に対して，何らかの基準によって相互間の類似度を与え，それを基に各個体を互いに似たもの同志のグループに分類しようとするものである。

クラスター分析の各手法の評価に関しては，様々の議論がなされているけれども，その必要とする時間及び空間的領域の面から考えても，G. N. Lance 及び W. T Williams の主張する組み合わせ的手法 (Combinatorial Method) が有効であるとされている。しかしながらデータに内在するクラスターの構造が linear-separable ではない場合に，これに含まれる手法が，どの程度有効であるのか，ということをここで述べる新しい手法との比較において検討することが本論文の目的である。

ここで特に linear-separable でない場合をとり上げる理由は，実際に観測及び測定によって得られるデータの構造 (特に因子構造) は，非常に複雑な場合が多いため，データの特性を表現

* 情報数理工学第一講座

する変量の組を座標と見なして，個々のもの（以下点と呼ぶことにする）を空間に配置した場合，単純に linear-separable な構造をもつことは少ない。従って実際に応用する立場から考えると，このような場合についての議論は重要なものであると考えられる。

2. 手法の概説

2.1 本論文における手法

与えられた，実数値をとる k 変量の N 個のデータ

$$x_{i1},\ x_{i2},\ \cdots,\ x_{ik} \qquad (i=1,2,\cdots,N)$$

に関して，N 個相互間の類似度を

$$s_{ij}=\cos\left\{\frac{d_{ij}-\min(d_{lm})}{\max(d_{lm})-\min(d_{lm})}\cdot\frac{\pi}{2}\right\} \qquad (i,j=1,2,\cdots,N)$$

により定義する。ここに d_{ij} はユークリッド距離を表わす。この s_{ij} からなる $(N\times N)$ の対称行列を $\boldsymbol{S}$ (similarity matrix) とするとき，$\boldsymbol{S}$ をもとに各点それぞれについて，自分自身を除いて最も類似度の大きい2点の対，および2番目に類似度の大きい2点の対の情報からなるリストを作成し，相互に2番目に大きい類似度の範囲で連結している2点の対の類似度を各点それぞれ自分自身を除く最大の類似度で置き換える。その結果得られる similarity 行列を $\bar{\boldsymbol{S}}$ とする。つぎに $\bar{\boldsymbol{S}}$ の各要素をそれぞれ3乗し，それらを対応する要素としてもつ行列を $\bar{\boldsymbol{S}}_1^{(3)}$ と表わす。さらに，この行列 $\bar{\boldsymbol{S}}_1^{(3)}$ に対してつぎのような変換をくり返し行なう。

$$\boldsymbol{D}_1\bar{\boldsymbol{S}}_1^{(3)\prime}\bar{\boldsymbol{S}}_1^{(3)}\boldsymbol{D}_1=\bar{\boldsymbol{S}}_2$$
$$\boldsymbol{D}_2\bar{\boldsymbol{S}}_2^{(3)\prime}\bar{\boldsymbol{S}}_2^{(3)}\boldsymbol{D}_2=\bar{\boldsymbol{S}}_3$$
$$\vdots$$

ここに $\boldsymbol{D}_i$ は $\bar{\boldsymbol{S}}_i^{(3)\prime}\bar{\boldsymbol{S}}_i^{(3)}$ の対角要素の平方根の逆数を対応する要素としてもつ対角行列である。この変換を $\bar{\boldsymbol{S}}_k$ のすべての要素が1又は0となるまでくり返す。このとき収束を早めるために $\bar{\boldsymbol{S}}_1^{(3)}$ の対角要素を各行の非対角要素の最大値でおき換えておくことは有効である。これは最も類似度の大きい2点は同一のクラスターに属するものと見なすことであり，クラスターを生成する上での妥当性を逸するものではないと考えられる。最終的に得られる行列において，その要素が互いに1であるもの同志を同一のクラスターを構成する要素と考える。この手法は直観的には，各点を空間に配置したとき，相対的に近いものは，さらに近づき，相対的に離れているものは，さらに離れ，最終的にクラスターの個数だけの点に収束するという特徴がある。

2.2 組み合わせ的手法

この手法においては，最初に各点を個々それぞれ1つのクラスターと考えて，互いに最も近いもの同志を順次結合し，新しいクラスターを生成していき，最終的に全体が1つのクラスターとなるまで続ける。このとき2つのクラスターを結合して新しいクラスターを生成したとき，このクラスターと他のクラスターとの距離を計算し直すことが必要である。これを結合する前の距離を用いて求めることができるところにこの手法の特徴がある。すなわち，データとして与えられる N 個相互間のユークリッド距離を $d_{ij}\ (i,j=1,2,\cdots,N)$ とし，ある段階において，クラスター p とクラスター q が結合されてクラスター r となったとき，このクラスター r と他のクラスター i との距離は

$$d_{ir}=a_pd_{ip}+a_qd_{iq}+bd_{pq}+c|d_{ip}-d_{iq}|$$

なる関係式で与えられる。このとき，上式のパラメータ a_p, a_q, b, c の与え方によって，それぞ

れつぎのような手法が考えられている。

(1) 最短距離法 (Nearest Neighbour Method)　各クラスター間の距離として，一方のクラスターに属する点と，他方のクラスターに属する点の間の距離で最小のものを，そのクラスター間の距離とする。すなわち

$$d_{ir}=\frac{1}{2}d_{ip}+\frac{1}{2}d_{iq}-\frac{1}{2}|d_{ip}-d_{iq}|.$$

(2) 最長距離法 (Furthest Neighbour Method)　クラスター間の距離として，相互のクラスターに含まれる点の間の距離の最大のものをとる。従ってこのときの距離は次式によって計算される。

$$d_{ir}=\frac{1}{2}d_{ip}+\frac{1}{2}d_{iq}+\frac{1}{2}|d_{ip}-d_{iq}|.$$

(3) Median Method　(1) と (2) の折衷的な方法であり，クラスター間の距離を

$$d_{ir}=\frac{1}{2}d_{ip}+\frac{1}{2}d_{iq}+bd_{pq}$$

として与える。ここで b は $-1/4\leqq b\leqq 0$ なる値をとるものとしているが，経験的に $-1/4$ が妥当であるということが，Lance 及び Williams によって示されている。ここでも $b=-1/4$ として計算する。

(4) 重心法 (Centroid Method)　クラスター p の中に n_p 個の点が含まれ，クラスター q の中に n_q 個の点が含まれているとき，これらを結合して得られるクラスター r と他のクラター i との距離を，クラスター p, q の重心間の距離と点の個数を考慮に入れ，つぎのように定義する。

$$d_{ir}^2=\frac{n_p}{n_r}d_{ip}^2+\frac{n_q}{n_r}d_{iq}^2-\frac{n_p}{n_r}\frac{n_q}{n_r}d_{pq}^2 \qquad (n_r=n_p+n_q)$$

(5) 群平均法 (Group-Average Method)　クラスター i とクラスター j にそれぞれ含まれる2点 i_α, j_β の距離を $e_{\alpha\beta}$ とするとき，この2つのクラスター間の距離を

$$d_{ij}^2=\frac{1}{n_in_j}\sum_\alpha\sum_\beta e_{\alpha\beta}^2$$

により定義する。ここに n_i, n_j はそれぞれクラスター i, j を構成する点の個数である。このとき，クラスター p とクラスター q が結合されたとき，他のクラスターとの距離は

$$d_{ir}^2=\frac{n_p}{n_r}d_{ip}^2+\frac{n_q}{n_r}d_{iq}^2$$

によって与えられる。

(6) Ward Method　クラスター i に関して，それを構成している各点の重心からの偏差平方和を I_i とする。クラスター p とクラスター q を結合してクラスター r とするとき，

$$\Delta I=I_r-I_p-I_q$$

なる ΔI が最小となるものを順次結合していく手法である。これを距離の関係で表現すると

$$d_{ir}^2=\frac{1}{n_i+n_r}\left[(n_i+n_p)d_{ip}^2+(n_i+n_q)d_{iq}^2-n_id_{pq}^2\right]$$

となる。この表現は W. Wishart によるものである。

3. 各手法による結果の比較検討

ここでは人工的に作成した2次元 (2変量) の linear-separable でないデータに関して，各手

法を適用し，その結果を検討する。次元を２次元に限定した理由は，説明をしやすくするためと，直観的にその結果の比較が容易となることからである。実際のデータに関して，もしそれが２次元で表現可能ならば，このような分析をする必要はないように思われる。すなわち，何らかの手段を用いてそれを図示し，人間の視覚で判断する方がより迅速であり正確である。しかしながら，理論的には，ここで用いている手法は次元とは無関係な議論であるため，２次元としてもその一般性を失なうものではない。

最初に図１のようなデータの配置が与えられた場合について考察する。この場合に図１のような２つのクラスターに分類できる手法は，本論文の手法，最短距離法，Ward 法の３つの手法である。本論文の手法では，互いに第二最近隣の範囲で連結されるもの同志に関しては，何らかの共通の要因をもつものとして，その類似度を大きくなるように修正した後，相対的に類似度の大きいものは互いに近づき，相対的に類似度の小さいものは互いに離れていくように変換をくり

表 1

データ 1 (図1)		データ 2 (図6)	
x	y	x	y
10.0	40.0	5.0	21.0
6.0	37.0	7.0	24.0
7.0	34.0	6.0	27.0
4.0	32.0	7.0	22.0
6.5	30.0	4.0	22.0
4.0	29.5	3.0	24.0
5.5	27.5	5.0	24.0
5.0	24.5	4.0	26.0
7.5	24.5	5.5	22.0
7.0	22.0	4.5	23.0
9.0	22.0	6.0	23.0
9.0	20.0	4.0	24.6
11.5	19.5	6.0	25.2
9.5	17.5	5.0	26.2
12.0	17.0	6.0	36.0
18.0	27.0	7.5	33.5
18.0	25.0	9.0	35.0
20.5	25.0	11.0	33.0
19.5	22.5	10.0	30.0
21.5	21.0	11.5	29.5
20.0	19.0	12.0	28.0
23.0	18.0	14.0	25.0
19.0	16.0	13.5	26.5
21.0	16.0	12.0	24.0
23.0	15.0	13.0	22.0
21.0	13.5	12.0	20.0
19.5	13.4	9.8	21.0
18.5	12.0	11.0	18.0
17.0	11.0	9.0	17.0
17.0	9.0	8.0	15.0
14.5	9.0	6.0	16.0
15.5	7.5	6.0	13.0
14.0	5.5		
13.0	7.5		
12.0	6.0		

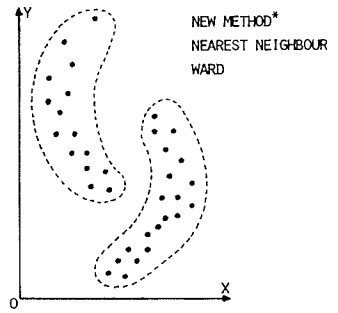

図 1

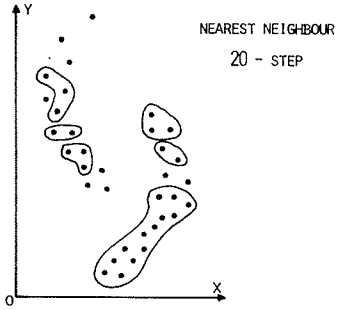

図 2-A

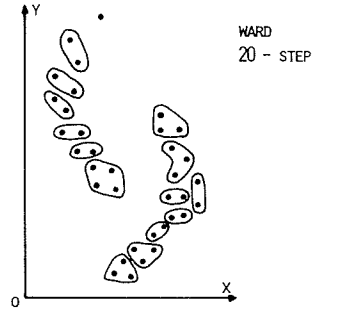

図 2-B

返すという特徴によってこのような分類が可能である。

最短距離法はクラスターが生成される度に，クラスター間の距離をその最短のものをとることによって，各クラスターが次第に近づいていく傾向にある。この意味において，本論文の手法と類似した性質をもっているが，それは一方的に近づくのみである。そのためこの手法では図1のように線状に分布している場合には比較的妥当な分類を得ることができる。この特徴は，Chain 効果と呼ばれている。一方この Chain 効果があるため，逆に弊害を伴うことが多い。さらに全体が一方的に近づいていく傾向にあるため，一般に分離感度が悪くなる。

Ward 法は空間の濃縮や拡散が生じないため，組み合わせ的手法の中にあっては，有効な手法とされているが，一般には，球状のクラスターが出来やすい。しかしながら図1のデータに関しては良好な結果が得られている。これは与えられたデータの2つのクラスターが，それぞれの重心に関して，比較的対称な形をしていることと，各々ほぼ一様な密度をもっていることが，妥当な結果が得られた理由と考えられる。

結果としては，この Ward 法と最短距離法は同じであるけれども，そのクラスター生成の過程を見ると両者の相違がより明確にわかる。図2-A, Bにそれぞれ最短距離法と Ward 法の20-step でのクラスターの出来る過程を示してある。最短距離法では chain 効果が生じていることがわかり，Ward 法においては，個々のクラスターがそれぞれ球状を保つように生成されていることがわかる。

これらに対して，最長距離法によるクラスターの構成が図3である。この手法は，クラスターが生成されるごとに，各クラスターが遠ざかっていくような傾向を示す。この性質により，点の分離感度はすぐれてはいるが，そのためこのデータのように線状に連らなっているクラスターを生成することは困難である。

図4に示してある結果は Median 法によるものであるが，これはクラスター間の距離として，最短距離法と最長距離法の中間的なものを考えているけれども，それは，両者の欠点を補なうというよりはむしろそれらの特徴が生かされず，性格があいまいなものになっている。

群平均法及び重心法については，その結果を図5に示す。これらの手法におけるクラスター間の距離の計算は，最短距離法や最長距離法と比較して，より妥当であると考えられる面があるけれども，クラスターの生成される特徴として，球状のものが生成される傾向がある。図5においてもその特徴を見ることができよう。従って球状のクラスターがいくつか存在する場合に，これらを分割する手法としては有効であるが，しかしそれも linear-separable でない場合には困難であることが多い。一方群平均法に比較して，Ward 法はクラスター内の点全体の平均からの偏差平方和を考えている点でよりすぐれているという立場がある。この例に関してもそれが示されている。

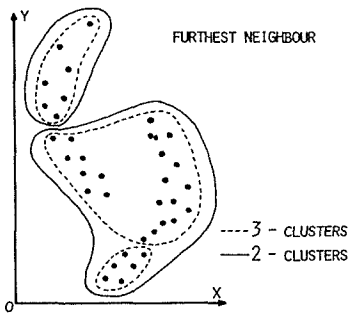

図 3

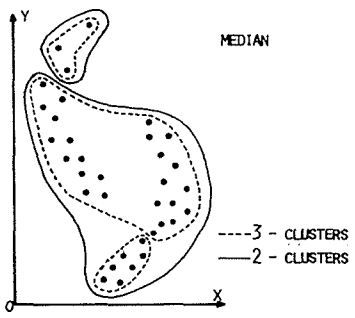

図 4

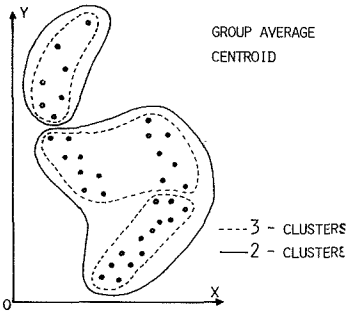

図 5

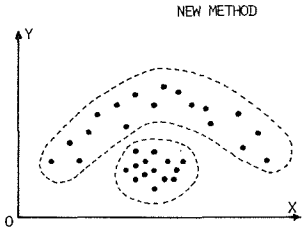

図 6

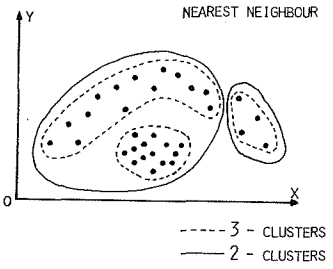

図 7

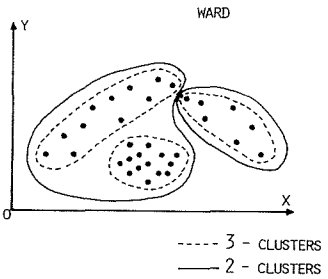

図 8

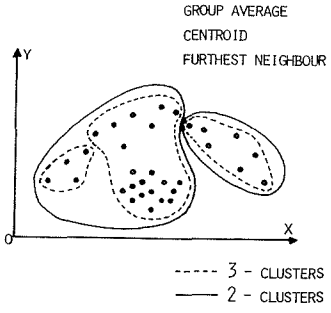

図 9

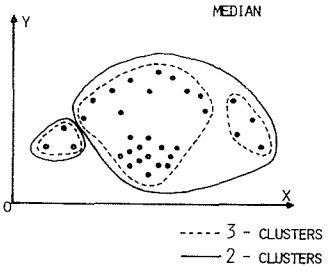

図 10

次に図６に示されるデータが与えられた場合の各手法により生成されるクラスターの相違について考察する。この場合には，もし図６に示されるような２つのクラスターが最も妥当であるとするならば，このようなクラスターが生成される手法は，ここでは本論文によって開発された手法のみである。この手法と他の手法の根本的な相違は，クラスターを生成する際に，点の配置を大域的に見ているか，局所的に見ているか，ということである。すなわち本論文の手法においては，最終的にクラスターが生成される過程の中で，常に個々の点と全体の点との関係を考慮に入れているのに対し，他の手法においては，最も接近している点及びクラスターのみに注目し，それらを順次結合していくのみである。そのため特にここで問題としている linear-separable でない場合においては，妥当な結果が得られないことが多い。

図７に示してあるのは，最短距離法による結果である。この場合には，３つのクラスターであると考えると，比較的良い結果であると見ることができるが，もしこれを２つのクラスターと見るならば図のような結果となることは，クラスター間の距離が，その最短のものをとることから理解されよう。

Ward 法においても3つのクラスターを考えるならば，比較的妥当なクラスターが得られることが図8に示されている。この手法は前に述べたように球状のクラスターを生成する場合には，他の組み合わせ的手法と比べて有効であると考えられるが，この例のように，球状のものとそうでないものが混在している場合には難点がある。このデータについて，2つのクラスターを考えるならば図9に示してある群平均法，重心法と一致した結果が得られている。

群平均法，重心法の結果は共に図9に示すとおり全く同じクラスターが生成される。これらは前に述べたように球状のクラスターが出来やすいという傾向がここにも現われている。さらにこれらと同様の結果が最長距離法によっても生成されていることは興味ある現象であるが，これはここで与えられたデータの性質に依存するものと考えられる。それは図6に示される2つのクラスターの密度が異なるため，最初に下にある密度の大きいクラスターが生成され，それと弓状に並ぶ点との相互関係により順次クラスターを生成していく際に，この手法は線状に並ぶものを強制的に分離する傾向をもつためこのような結果が得られるのである。

Median 法による結果は図10に示されているとおりである。この手法はクラスターが生成される度に空間が濃縮されていく傾向にはあるが，それも最短距離法ほど極端ではなく，比較的平均距離に近いため，クラスター間の関係も平均化され全体としての分離感度は悪く，全体が1つにまとまる傾向を示す。従ってこの手法は他の手法と比較して，実際にはそれほど有効な手法であるとは考えられない。

4. 結　　論

以上述べた各手法において，組み合わせ的手法に含まれるものについては，ここで問題とした linear-separable でない場合に，そのデータの配置が線状をなしているものについては，最短距離法が比較的有効であり，また Ward 法は全体としてすぐれた手法であることは，本論文の例からもいえることである。しかしながらこれらの手法はいずれも，どの段階を最終的なクラスターと考えるのか，ということが明確ではないところに問題がある。ある意味においては，このことはデータを解釈する上で利点はあるけれども，都合の良い結果だけをを抽出する危険性がある。また得られた結果が真にクラスターと考えて良いものかどうかの判定をすることがこの手法のみでは不可能である。この意味において本論文により，クラスターとは相互に共通の性格をもつものの集まりとして，それを一意に定める手法の開発が行なわれたのである。さらに本論文において，線形分離不可能なデータに関しても，この手法の応用が可能であることを他の手法との比較において示したのである。

実際に分析の対象として与えられるデータは，ここで示した以上に複雑な構造をもったものが多いのであり，この新しい手法によりすべてを解決できるわけではないが，現在まで考えられている手法よりは有効な点があると思われる。しかしこの手法においても分離可能な限界がある。例えば2つのクラスターの最短距離が各クラスター内の第二最近隣よりも小さくなる程度接近したものについては分離不可能である。

尚最後に本論文のデータ作成に当たり，その計算に御協力いただいた情報数理工学第一講座大学院生，宮腰政明氏に感謝の意を表するものである。

参 考 文 献

1) Y. Sato: Master thesis, Division of Information Engineering. Hokkaido Univ. (1975).
2) 矢島敬二，王　碩夫: オペレーションズリサーチ (1971).

3) R. Sibson: Comp. Jour. **14** (1970) 156.
4) N. Jardine and R. Sibson: Comp. Jour. **14** (1971) 404.
5) E. H. Rrspini: Information and Control **15** (1972) 22.
6) G. N. Lance and W. T. Williams: Comp. Jour. **12** (1969) 60.
7) W. T. Williams, G. N. Lance, M. B. Dale and H. T. Clifford: Comp. Jour. **14** (1971) 162.
8) G. N. Lance and W. T. Williams: Comp. Jour. **14** (1971) 154.
9) G. Nagy: Proc. IEEE. **56** (1968) 836.
10) M. G. Kendall: Frontiers of Pattern Recognition (Watanabe ed.), Academic Press (1972) 291.
11) M. R. Anderberg: Cluster Analysis for Applications (1973) Academic Press.