クラスター化法を用いた分類結果の構造

竹内 光悦　立教大学社会学部

E-mail: akitake@rikkyo.ac.jp

1　はじめに

情報化時代の到来により, 様々な分野で多種多様なデータが容易に入手することが可能となり, これらを適切に分類する方法の必要性が増している. 分類法の中でもクラスター分析は様々な種類のデータに対応でき, またその中でも一連の手法の総称である凝縮型階層的クラスタリング法 (Agglomerateve Hierarchical Clustering Algorithm. 以下, AHCA) は分類結果を樹形図で表すことができ, 直感的で分かりやすい解釈が可能なことから多くの分野で利用されている. しかしながら一連の手法といいながらも, 同一のデータを分析しても選択する手法により大きく異なった分類結果が得られることが多々見られる. このような現象が起きながら手法選択に関して, 現在までにいくつか基準が提案されているがあまり知られておらず, 実際すべてのデータや分析意図に相応しい手法が存在するわけではなく, 分析者の主観において恣意的に選択されている点は問題といえよう. これまでに提案されよく知られている指標では, 例えば, Cophenetic correration coefficient (Sokal and Rohlf, 1962), Sum of squares (Hartigan, 1967), Minkowski metrics (Jardine and Sibson, 1971) 等があげられる. 一方, クラスターを形成している対象の散らばり具合を基準にした分類の良さを表す概念が Rubin (1967) によって提案された. 彼は同一のクラスターに属する 2 つの対象間の距離の最大値が, 異なるクラスターに属する 2 つの対象間の距離の最小値より大きい, すなわち

$$\max_{I} \max_{p,q \in C_I} d_{pq} < \min_{\substack{I,J,\\ I \neq J}} \min_{p \in C_I, q \in C_J} d_{pq} \tag{1}$$

を与えられた対象を表すデータが満たすとき, このデータを well-structured (L-group) データと名付けた. その後, Fisher and Van Ness (1971) は Rubin が提案した概念を用いた許容性として well-stuructured (L-group) admissible を提案した. 彼等は任意の well-structured (L-group) データが, L 個のクラスターに分けられたとき, 対象を表すデータが上式を満たすならば, 用いた分類法は well-structured (L-group) admissible であると定義した. Chen and Van Ness (1996) は well-structured (L-group) admissible とその他の許容性の関係を述べ, 加えて well-structured (L-group) admissible と更新式のパラメータとの関係を与える.

本報告では, Rubin (1967) によって提案された well-structured の概念に基づき, 階層的クラスタリング法 (Agglomerative hierarchical clustering algorithm, 以下 AHCA) による各結合結果に対して “構造度” と呼ばれる分類結果のある種の “良さ” を表す尺度を提案する. 加えて, 構造度に基づく分類法の許容性及び新たな分類法, また直感的に構造度を解釈するための視覚的な表現法についても提案する.

なお, 本報告においては, 対象 p と対象 q 間の非類似度を d_{pq} で表し, 分類すべき対象の個数を N とし, m $(1 \leq m < N)$ 段階目の結合時のクラスター I を $C_I(m)$ で表す. $n_I(m)$ を $C_I(m)$ に

属する対象の個数とし, $n_I = 1$ のとき ${}_{n_I}C_2 = 1$ であるとする. また, 同一数式内で I, J が同時に用いられている場合, $I \neq J$ が仮定されているとする.

2　構造を基準とした指標

本節では, 同一のクラスターに属する対象の散らばりの程度を表す “クラスター内散布 W” と異なるクラスターに属する対象のそれを表す “クラスター間散布 B” をそれぞれいくつか定義し, それらを用いて, Rubin (1967) の well-structured の概念を拡張する構造度を定義する. また, その性質についても述べる.

定義 2.1. 第 m $(< N-1)$ 段階目の結合における構造度 (structured ratio) を以下の式で定義する.

$$SR_h(m) = W_h(m)/B_h(m) \tag{2}$$

ここで W_h と B_h はそれぞれ m 段階目におけるクラスター内散布とクラスター間散布であり, 以下の式によって定義される.

$$W_1(m) = \max_I \max_{p,q \in C_I(m)} d_{pq}, B_1(m) = \min_{I,J} \min_{\substack{p \in C_I(m) \\ q \in C_J(m)}} d_{pq} \tag{3}$$

$$W_2(m) = \sum_I \left(\max_{p,q \in C_I(m)} d_{pq} \right) \Big/ (N-m), B_2(m) = \sum_{I,J} \left(\min_{\substack{p \in C_I(m) \\ q \in C_J(m)}} d_{pq} \right) \Big/ {}_{N-m}C_2 \tag{4}$$

$$W_3(m) = \max_I \left(\sum_{p,q \in C_I(m)} d_{pq} / {}_{n_I}C_2 \right), B_3(m) = \min_{I,J} \left(\sum_{\substack{p \in C_I(m) \\ q \in C_J(m)}} d_{pq}/n_I n_J \right) \tag{5}$$

$$W_4(m) = \sum_I \left(\sum_{p,q \in C_I(m)} d_{pq} / {}_{n_I}C_2 \right) \Big/ (N-m),$$

$$B_4(m) = \sum_{I,J} \left(\sum_{\substack{p \in C_I(m) \\ q \in C_J(m)}} d_{pq}/n_I n_J \right) \Big/ {}_{N-m}C_2 \tag{6}$$

$$W_5(m) = \sum_I \left(\sum_{p,q \in C_I(m)} d_{pq} \right) \Big/ \sum_I {}_{n_I}C_2, B_5(m) = \sum_{I,J} \left(\sum_{\substack{p \in C_I(m) \\ q \in C_J(m)}} d_{pq} \right) \Big/ \sum_{I,J} n_I n_J \tag{7}$$

構造度は, クラスター内散布とクラスター間散布の比であり, その値が小さい程, ある意味で良い分類であるといえよう. ここで多くの散布を挙げているのは, 散布の選び方によって構造度の値に大きな変化があり, 1 種類の散布に基づく構造度で解釈するのは問題があると考えたためである. とはいえ, それぞれの散布を用いて求められた構造度の値をいかに解釈するのかとう問題は残っている.

構造度は各結合に対して定義されたが, 分類全体に対しては以下のような平均構造度を定義することができる.

定義 2.2. 平均構造度 (total structured ratio) を以下の式で定義する.

$$TSR_h = \sum_{m=1}^{N-L} SR_h(m)/(N-L) \tag{8}$$

ここで L $(1 < L < N)$ は解析の結果選択されるクラスターの個数である.

最終的な分類結果を測るのであれば $SR_h(N-L)$ を用いればよいが, 次節で論じられるような, 分類法の (構造度の意味での) 良さを測るためにはこのような指標も必要であると考えている.

3 構造度を用いた許容性

本節では, 構造度, 平均構造度を用いて, 分類法の許容性を提案する. Fisher and Van Ness (1971) が提案した well-structured (L-group) admissible であるための必要十分条件は構造度を用いると, 任意の well-structured (L-group) データに対して

$$SR_1(N-m) < 1 \tag{9}$$

が成立することである. この概念を特別な場合として含む許容性が以下のように定義される.

定義 3.1. m $(= N-L)$ 段階目において, 対象が解析の結果 L $(1 < L < N)$ 個のクラスターに分割されたとする. このとき,

$$SR_h(m) < \zeta \tag{10}$$

が成立するならば, AHCA は SR_h に関して ζ 構造許容的 (L-group) であるという.

定義 3.2. 対象が解析の結果 L $(1 < L < N)$ 個のクラスターに分割されたとする. このとき, 任意の m $(1 \leq m \leq N-L)$ に対して,

$$SR_h(m) < \zeta \tag{11}$$

が成立するならば, AHCA は SR_h に関して ζ 構造許容的 (perfect) であるという.

定義 3.3. 対象が解析の結果 L $(1 < L < N)$ 個のクラスターに分割されたとする. このとき,

$$TSR_h < \zeta \tag{12}$$

が成立するとき, AHCA は SR_h に関して ζ 平均構造許容的であるという.

定義から明かのように, ζ 構造許容的 (L-group) は L 個に分類されたときのみを問題にしている. そのため, ζ 構造許容的 (perfect) より緩やかな条件である. 実際, L が小さいときに AHCA が ζ 構造許容的 (perfect) であるためには, 相当大きな ζ を取る必要がある. 以下に, いくつかの性質をあげておく.

性質 3.1. AHCA が ζ 構造許容的 (perfect) であれば, その AHCA は ζ 構造許容的 (L-group) である.

性質 3.2. AHCA が ζ 構造許容的 (perfect) であれば, その AHCA は ζ 平均構造許容的である.

性質 3.3. AHCA が SR_h に関して 1 構造許容的 (L-group) であるとき, その AHCA は well-structured (L-group) admissible である.

4　構造度に基づく分類法

本節では構造度を用いて各結合段階で最小の構造度をもつ AHCA を提案する.

定義 4.1. 第 1 段階では, 最小の対象間距離をもつ対象が結合し, その後, m (≥ 2) 段階における構造度が最小となるように, $m-1$ 段階のクラスターを結合させる AHCA を MSR_h 法 (SR_h による Minimizing structured ratio 法) という.

ここでは, 理論的にはこのような分類法が提案できることを述べるにとどめたい. この分類法により分類を行うにはかなりの計算を必要とするのは明らかであり, 現在その実用性を確認中である.

5　構造度の視覚的表現

通常, AHCA を用いた分析結果は, 樹形図を用いて表現されている. 樹形図では, 一方の軸でクラスター同士が結合する際の距離, 他方の軸で対象を表すのが普通である. 構造度と分類結果の同時表現を考えるとき, 樹形図内のクラスター同士が結合した箇所に構造度の値を書き込むことや結合距離の代わりに構造度を用いることが考えられるが, 直観的に見やすいとは言い難い. そこで本節では構造度を視覚的に表現し, 分類結果と共に表示することで, 分類の特徴をより明確に表現することを提案する.

クラスター同士の m $(1 \leq m < N-1)$ 段階目の結合を図 5 の二分木で表現する. ここで, 各結合段階における $\theta(m), l(m)$ の値をそれぞれ

$$\theta(m) = 2\arctan(SR_h(m)), \quad l(m) = B_h(m)$$

とする.

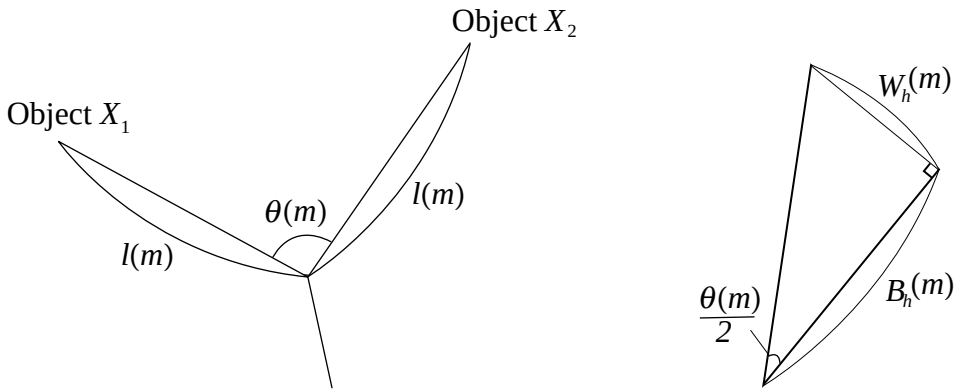
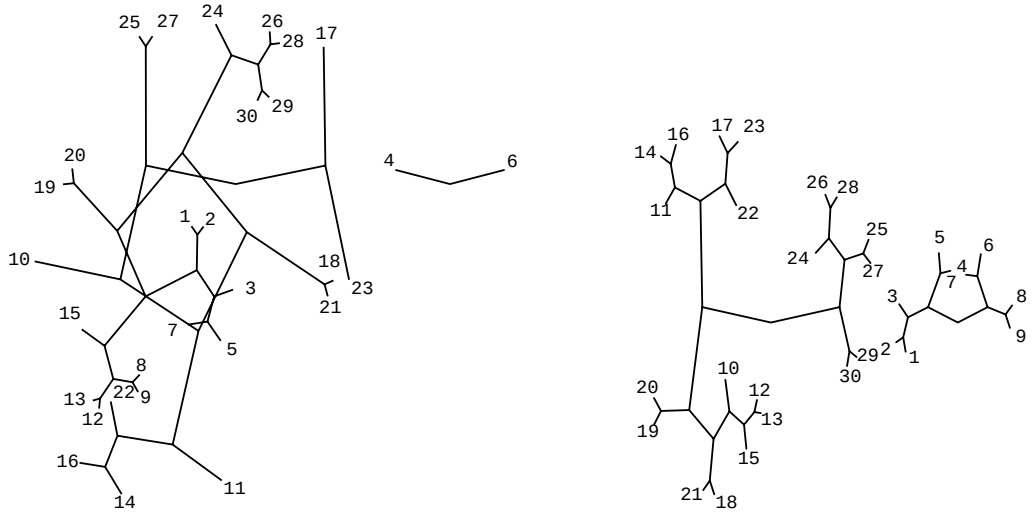


図 1: m 段階目における結合の表現

ここで, $\theta(m)$, $SR_h(m)$ の関係は $\tan(\theta(m)/2) = SR_h(m)$ なので, $SR_h(m)$ の値が小さいとき, すなわち, 結合の際に同じクラスターに属する対象の散らばりが少なく, 異なるクラスターに属する対象の散らばりが大きいとき, 二分木の狭角が狭くなることを示している. このとき 2 つの対象は近くに表示される. 逆に, $SR_h(m)$ の値が大きい場合は狭角が広がり, 2 つの対象は離れて表示される.

なお, m 段階目の作図における辺の長さを m 段階目の結合距離とすること (すなわち $l(m) = d_{(IJ)K}$) も考えられるが, ここでは更新式で表せない分類法 (例えば MSR_h 法) に対しても同様に定義することを考えているため上記のように提案した.

図 5, 5 は 30 個の対象からなる人工データをそれぞれ最短距離法と最長距離法を用いて分析したときの結果を上記の表現で表したものである. 最短距離法を用いた分析結果 (図 5) では, 各結合段階の構造度を表す狭角が大きくなり, 全体を通して, 結合段階の構造度が大きいことが読み取れる. 一方, 最長距離法を用いた分析結果 (図 5) では, 狭角が小さく, 結合段階の構造度が小さいことが読み取れる.



最短距離法　　最長距離法

図 2: 構造度を用いた分類結果の視覚的表現

6　おわりに

構造度は手法のみに依存するのではなく, 与えられたデータと手法の両方に依存し, 分析結果を評価する指標である. しかしながら, その性質等未解明な点も多く, 課題も多く残っている. 今後, 数値実験等を用いて解明する予定である.

一方視覚的表現においては, 構造度と分類結果を同時に表現することにより, 各結合段階における構造度の大きさだけではなく, クラスター内外の対象の散らばり具合が視覚的に判断でき, 結合したクラスターの組も容易に読み取ることが可能になったと考えている.

参考文献

[1] Fisher, L. and Van Ness, J. (1971) Admissible clustering procedures, *Biometrika*, **58**, 91–104.

[2] Gordon, A. D. (1996) Hierarchical classification. In *clustering and classification*. World Scientific, New Jersey.

[3] Hartigan, J. A. (1967) Representation of similartity matrices by trees, *Journal of the American Statistical Association*, **62**, 1140–1158.

[4] Jardine, N. and Sibson, R. (1971) *Mathematical taxonomy*, London, Wiley.

[5] Lance, G. N. and Williams, W. T. (1967) A general theory of classificatory sorting strategies: 1. hierarchical systems, *The Computer Journal*, **9**, 373–380.

[6] Rubin, J. (1967) Optimal classification into groups: an approach for solving the taxonomy problem, *Journal of Theoretical Biology*, **15**, 103–144.

[7] Sokal, R. R. and Rohlf, F. J. (1962) The comparison of dendrogtams by objextice methods, *Taxon*, **11**, 33–40.

[8] 竹内光悦, 宿久洋, 稲田浩一 (1999) 凝縮型階層分類法による分類結果の構造について, 日本行動計量学会第 *27* 回大会発表論文抄録集, 305–308.

[9] 竹内光悦, 宿久洋, 稲田浩一 (2000) 凝縮型階層的分類結果のシミュレーションによる評価と視覚的表現, 日本行動計量学会第 *28* 回大会発表論文抄録集, 225–228.

[10] Takeuchi, A., Yadohisa, H., and Inada, K. (2000) Measuring the structure of the agglomerative hierarchical clustering, *Proceedings of The International Conference on Measurement and Multicatiate Analysis*, **1**, 18–20.

[11] Takeuchi, A., Yadohisa, H., and Inada, K. (2001) A simulated study of measures for clustering, Bulletin of *the International Statistical Institute, 53rd Session Contributed Papers*, **1**, 179–180.