計算機統計学
第15巻・第 2 号：2002
309 – 316

非対称凝縮型階層的クラスター分析法の定式化とその解析結果の視覚的表示法の提案

宿久　洋*

要　旨

非対称凝縮型階層的クラスター分析法のアルゴリズムを統一的に取り扱うための拡張更新式を提案する．この更新式は Lance and Williams (1967) の更新式を非対称データに対応するように拡張したものであり，これにより従来は提案者によって様々な形で定義されていた非対称凝縮型階層的クラスター分析法を統一的に表現することが可能となる．合わせて，これらの手法による解析結果を視覚的に表現するための拡張デンドログラムも提案する．このデンドログラムは，既に 岡田, 岩本 (1995), Okada & Iwamoto (1996) により同じ目的のために提案されているデンドログラムが表現可能な情報をすべて表すことが可能であり，加えていくつかの新たな情報も表現することができる．

1.　はじめに

クラスター分析において用いられる(非)類似度データ(単相 2 元データ)は一般に非対称なものである．すなわち，対象 A からみた対象 B の(非) 類似度が逆の場合のそれと一般に異なっている．しかしながら，実際に分析を行う際にはこのような非対称性は無視され，平均化する等の適当な対称化を行った後，対称クラスター分析法によって分析したり，データ行列の上側三角部分と下側三角部分が別々のものを表している (2 相 2 元データ) と考え，分析を行ったりするということが一般的に行われている (Arabie et al., 1988; DeSarbo & De soete, 1984)．これに対し，通常は無視されている非対称性には何らかの本質的な意味があり，非対称性を考慮した解析を行うべきであるという考えに基づき，いくつかの分析手法が提案されている．Hubert (1973) では，与えられた非対称類似度行列を最初に対称化し，その後，分析を行う Min and max clustering が提案されている．藤原 (1980) は，Hubert (1973) を拡張し，最初に対称化を行わず，非対称性を保ったままで分析を行う手法を提案している．これら 2 つの論文における提案は対称クラスター分析法の場合の最短距離法及び最長距離法を非対称に拡張したものであった．これに対し，岡田, 岩本 (1995), Okada & Iwamoto (1996) では，対称の場合の加重平均法を非対称に拡張したものを提案している．この他にも，Brossier, G. (1982), Ozawa (1983), DeSrbo et al. (1990) などで異なるアプローチによる提案がなされている．

これらの論文において，各手法は，(i) 結合するクラスター (対象) を選択する，(ii) 新たに結合されたクラスターと他のクラスター(対象)との(非)類似度を求める．という 2 つのプロセスを如何に定めるかによって個別に定義されていた．

本論文では，非対称クラスター分析法の中でも凝縮型階層的なもの (Asymmetric Agglomerative Hierarchical Clustering Algorthm, 以下,

* 鹿児島大学理学部 〒890-0065 鹿児島市郡元1－21－35 (Tel. 099-285-8040),　E-mail: yado@sci.kagoshima-u.ac.jp
本論文は，計算機統計学会15周年記念誌への寄稿論文である．計算機統計学会奨励賞受賞者として，その後の研究について，最近，興味を持っているテーマについての論文という形で書かせていただいた．このような機会をいただき大変光栄に思っている．関係各位に謝意を表したい．

AAHCA と略す) に着目し, これらの手法を, 対称クラスター分析法における Lance & Williams (1967) の更新式のように統一的に扱うための拡張更新式を提案する. 合わせて, AAHCA の解析結果の表示のための拡張デンドログラムの提案も行う.

2. 既存の AAHCA

この分野の先駆的な研究は Hubert (1973) によるものであるが, ここでは, その後, 藤原 (1980) によって, Hubert (1973) の手法を含む形で提案された表記に基づいて紹介する. 彼の提案している手法は非対称データを与えられた際, 最初に対称化するものとしないのものの 2 つに分けられる.

対称化する場合は, 前述のプロセス (i) として, 与えられた非対称類似度行列 $\boldsymbol{S} = [s_{ij}]$ を

$$(\mathrm{A}) s^*_{ij} = s^*_{ji} = \max(s_{ij}, s_{ji})$$
$$(\mathrm{B}) s^*_{ij} = s^*_{ji} = \min(s_{ij}, s_{ji})$$

のいずれかで対称類似度行列 $\boldsymbol{S}^* = [s^*_{ij}]$ に変換した後, $s^*_{pq} = \max_{i<j}(s^*_{ij})$ を満たす対象 p 及び q を含むクラスター C_I 及び C_J を選択し, プロセス (ii) として, C_I と C_J を結合してできるクラスター C_{IJ} と他のクラスター C_K の類似度を

$$(\mathrm{a}) s^*_{ro} = s^*_{or} = \max(s_{po}, s_{qo})$$
$$(\mathrm{b}) s^*_{ro} = s^*_{or} = \min(s_{po}, s_{qo})$$
$$(\text{但し}, r \in C_{IJ}, o \in C_K)$$

のいずれかで定めるものである.

対称化しない場合は, プロセス (i) として,

$$(\mathrm{C}) \max(s_{pq}, s_{qp}) = \max_{i<j}(\max(s_{ij}, s_{ji}))$$
$$(\mathrm{D}) \min(s_{pq}, s_{qp}) = \max_{i<j}(\min(s_{ij}, s_{ji}))$$

のいずれかを満たす対象 p 及び q を含むクラスター C_I 及び C_J を選択し, プロセス (ii) として C_I と C_J を結合してできるクラスター C_{IJ} と他のクラスター C_K の類似度を

$$(\mathrm{c}) s^*_{ro} = \max(s_{po}, s_{qo}),\ s^*_{or} = \max(s_{op}, s_{oq})$$
$$(\mathrm{d}) s^*_{ro} = \min(s_{po}, s_{qo}),\ s^*_{or} = \min(s_{op}, s_{oq})$$
$$(\text{但し}, r \in C_{IJ}, o \in C_K)$$

のいずれかで定めている. このうち Hubert (1973) で提案されたものは, ステップ (i) が A, ステップ (ii) が a 及び b の 2 つの場合とステップ (i) が B, ステップ (ii) が b の場合の 3 とおりである. なお, これらの 2 つの論文においては手法の提案のみならず, グラフ理論を用いた非対称性の表現についても提案されている.

Okada & Iwamoto (1995) は, プロセス (i) は最初に対称化しない場合と同様で, プロセス (ii) を

$$(\mathrm{e}) s^*_{ro} = (s_{po}, s_{qo})/2,\ s^*_{or} = (s_{op}, s_{oq})/2$$
$$(\text{但し}, r \in C_{IJ}, o \in C_K)$$

のいずれかで定める手法を提案している. なお, この論文では, クラスター化結果の樹形図による表現を非対称な場合に拡張したもの, 及び, 自己類似度 (類似度行列の対角成分) も考慮した表現法についても提案されている.

3. 拡張更新式と AAHCA

前章における各手法の記述方式は藤原 (1980) に準じて統一して行ったが, 実際はそれぞれの著者毎に異なった記述方式がとられている. もちろん, 手法を定義するにはそれで十分であるが, いくつかの手法の解析結果を比較したり, 計算機にインプリメントする際には些か不便である. そこで, 対称凝縮型階層的クラスター化化法における Lance & Williams (1973) の更新式に対応するものを非対称の場合にも定義し, より統一的な形で各手法を取り扱うことを提案する.

定義 1：拡張更新式

クラスター C_I とクラスター C_J が結合してできたクラスター C_{IJ} からみた他のクラスター C_K との更新距離 $d_{(IJ)K}$, 及び他のクラスターC_K からみたクラスター C_{IJ} との更新距離 $d_{K(IJ)}$ をそれぞれ以下のように定義する.

$$\begin{aligned} d_{(IJ)K} &= \alpha_I^1 f^1(d_{IK}, d_{KI}) + \alpha_J^1 f^1(d_{JK}, d_{KJ}) \\ &\quad +\beta^1 g^1(d_{IJ}, d_{JI}) \\ &\quad +\gamma^1 |f^1(d_{IK}, d_{KI}) - f^1(d_{JK}, d_{KJ})| \\ d_{K(IJ)} &= \alpha_I^2 f^2(d_{IK}, d_{KI}) + \alpha_J^2 f^2(d_{JK}, d_{KJ}) \\ &\quad +\beta^2 g^2(d_{IJ}, d_{JI}) \\ &\quad +\gamma^2 |f^2(d_{IK}, d_{KI}) - f^2(d_{JK}, d_{KJ})| \end{aligned}$$

ここで d_{IJ} はクラスター C_I とクラスター C_J の非類似度であり, α_I^1, α_J^1, α_I^2, α_J^2, β^1, β^2, γ^1, γ^2 は解析前に決定されている定数, f^1, f^2, g^1, g^2 は解析前に決定されている 2 つの非類似度を引数とする関数である.

この更新式を使うことにより, 対称な場合と同様, m 段階のクラスター (対象) 間距離から $m+1$ 段階のそれらを決定することができる. すなわち, 上記の定数及び関数を定めることはクラスター分析法を決定することに対応し, それぞれの場合に対して単一のクラスター化結果を与えることになる.

拡張更新式においても通常の更新式と同様, α_I^1 等のパラメータはクラスターの結合を制御するものである. 拡張更新式が 2 つの式からなるのは, 類似度行列の上側三角部分と下側三角部分を別々に更新しているからであるのは言うまでもない. また, 関数 f^1, f^2, g^1, g^2 は非対称性を制御するものである. 拡張更新式内の各クラスター間距離の関係を図 1 に示す.

次に, 拡張更新式を用いて AAHCA を定義する.

定義 2：拡張更新式を用いた AAHCA

非対称非類似度行列 $\boldsymbol{X} = (d_{ST})$ が与えられているとする.

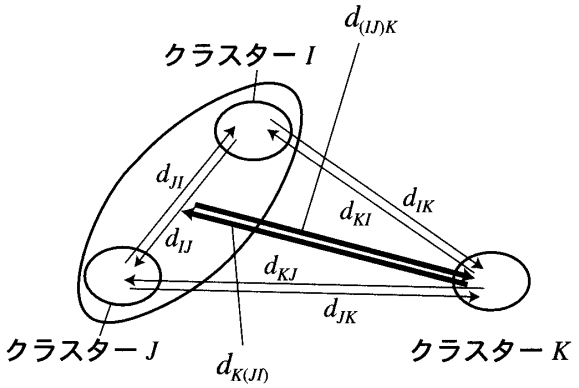

図 1: 非対称なクラスター間非類似度

プロセス (i): 以下を満たすクラスター C_I 及び C_J を選択し, クラスターC_{IJ} を作る.

$$d_{IJ} = \min_{S<T} D(d_{ST}, d_{TS}) \qquad (3.1)$$

ここで, D はデータ行列の対応する 2 つの非対角要素を引数とする結合の基準であり, max, min, mean, など様々なものが考えられる.

プロセス (ii): 拡張更新式を用いて, $d_{(IJ)K}$ 及び $d_{K(IJ)}$ を更新する.

プロセス (i), (ii) を 1 つのクラスターになるまで繰り返す.

このアルゴリズムにおいて, C_S, C_T 及び d_{ST}, d_{TS} は任意のクラスター及びそれらの間の非類似度を表している. また, C_{IJ}, C_K その段階で結合する特定のクラスターを表している. また, 対称の場合の更新式と同様に, 一般性を失うことなく, C_K は単一の対象のみからなるクラスター, 又は, C_{IJ} より以前に結合したクラスターであること, 及び, $D(d_{IJ}, d_{JI}) \leq D(d_{IK}, d_{KI}) < D(d_{JK}, d_{KJ})$ が仮定されている.

表 1 に, 第 2 章において取り上げた各手法を拡張更新式を用いて表現する. 以下の(A - a)等の表記が前章で述べた 2 つのプロセスでの選択に対応している. すなわち, (A - a)は(i)でAを選

表 1: 拡張更新式のパラメータと既存の AAHCA

手法	規準 D	$\alpha_i^1(=\alpha_i^2)$	$\alpha_j^1(=\alpha_j^2)$	$\beta^1(=\beta^2)$	$\gamma^1(=\gamma^2)$	$f^1(=g^1)$	$f^2(=g^2)$
(A–a)	min	1/2	1/2	0	−1/2	$\min(x,y)$	$\min(x,y)$
(A–b)	min	1/2	1/2	0	1/2	$\min(x,y)$	$\min(x,y)$
(B–a)	max	1/2	1/2	0	−1/2	$\max(x,y)$	$\max(x,y)$
(B–b)	max	1/2	1/2	0	1/2	$\max(x,y)$	$\max(x,y)$
(C–c)	min	1/2	1/2	0	−1/2	x	y
(C–d)	min	1/2	1/2	0	1/2	x	y
(C–e)	min	1/2	1/2	0	0	x	y
(D–c)	max	1/2	1/2	0	−1/2	x	y
(D–d)	max	1/2	1/2	0	1/2	x	y
(D–e)	max	1/2	1/2	0	0	x	y

択し, (ii) で a を選択することに対応している.

藤原 (1980) でも指摘されていることであるが, (A - a) と (C - c) 及び (B - b) と (D - d) は同等な手法である. また, Algorithm I, II において, (A - a), (A - b), (B - b) の結果は Lubert (1973) と一致するが, 事前の対称化は行っていないことを注意しておく.

拡張更新式のパラメータ, 関数 f, g, 結合の基準 D を定めることにより, 多くの非対称クラスター分析法を定義することができる. Hubert (1973), 藤沢 (1980) で提案されているのは, 対称の場合の最短距離法, 最長距離法に対応するものであり, Okada & Iwamoto (1996) のそれは対称の場合の加重平均法に対応するものであった. 拡張更新式を用いることにより, その他の一般的な手法も非対称な場合に拡張することができるのは言うまでもない.

4. 解析結果の視覚的表示

通常のデンドログラムでは, AAHCA の解析結果を (非対称性も含めて) 表示することはできない. そこで, デンドログラムではクラスターの結合を表現するにとどめ, 非対称性を表現するための何らかのグラフを併記することが行われている.

これに対して, 対象 p と対象 q の間の非対称な類似度 S_{pq} 及び S_{qp} の表現法として, 岡田, 岩本 (1995) では, 非対称性の相対的な大きさを表現するデンドログラムを, また, Okada & Iwamoto (1996) では, 非対称性を結合の方向と考えることにより, 結合と非対称性を同時に表示するデンドログラムを提案している (図 2, 図 3 参照).

ここでは, 結合時に選択されたクラスター間距離 (結合の規準) $D(d_{(IJ)K}, d_{K(IJ)})$, それらのクラスター間の最大距離 $\max(d_{(IJ)K}, d_{K(IJ)})$ および最小距離 $\min(d_{(IJ)K}, d_{K(IJ)})$ を表現する拡張デンドログラムを提案する (図 4 参照).

これにより, 結合の際に選択した距離が非対称性の意味でどのような位置にあるかを読み取ることができる. 最後に, 拡張デンドログラムは岡田, 岩本 (1995) 及び Okada & Iwamoto (1996) のデンドログラムが持つ情報の全てを含んでいることを注意しておく.

5. 数値例

ここでは, Stigler (1994) による統計関係雑誌間の引用関係のデータをを既存の非対称最短距離法の 4 種類の方法の解析結果を拡張デンドログラムを用いて表現する.

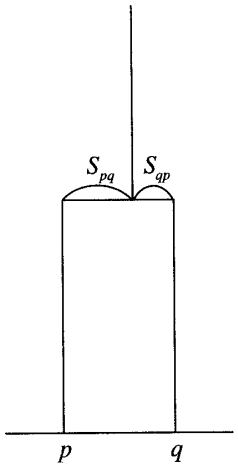

図 2: 岡田, 岩本 (1995)

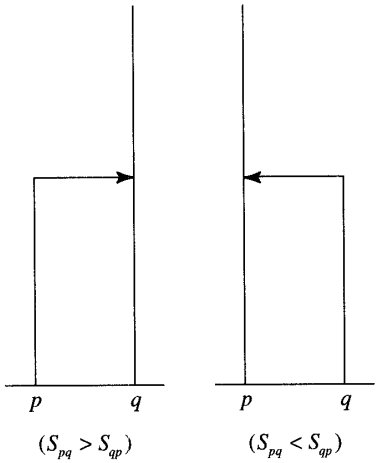

図 3: Okada and Iwamoto (1996)

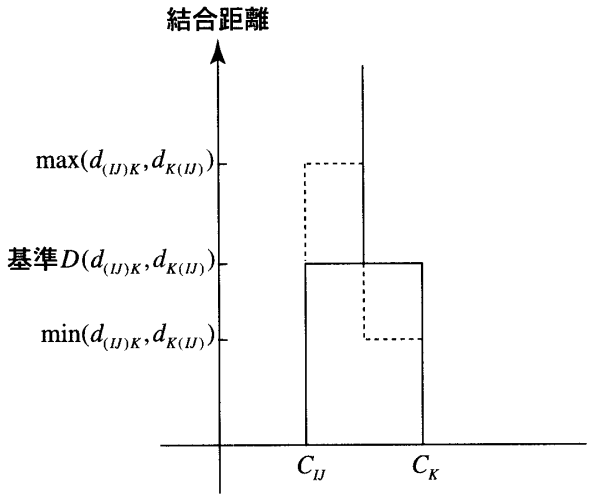

図 4: 拡張デンドログラム

表 2 は1987年から1989年までの 8 つの雑誌間の引用数のデータである. 各雑誌は "The Annals of statistics", "Biometrics", "Biometrika", "Communications in Statistics", "The Journal of the American Statistical Association", "The Journal of Royal Statistical Society (Series B)", "The Journal of Royal Statistical Society (Series C)", "Technometrics" であり, 表ではそれぞれ, "AnnSt", "Biocs", "Bioka", "ComSt", "JASA", "JRSSB", JRSSC", "Tech" と略されている.

例えば, 42 本の Biometrics に掲載された論文が Annals of Statistics に引用されていることを表し, 逆に155 本の Annals of Statistics の論文が Biometrics に引用されていることを示している.

図 5 ～図 8 はそれぞれ, 手法 (A – a), 手法 (B – a), 手法 (C – c), 手法 (D – c) での解析結果を表している. これらはすべて最短距離法に基づくものであるが, 初めの 2 つの結果には非対称性は見られない. これは, この 2 つの手法が事前に対称化を行う手法と一致するものだからである. 残りの 2 つの手法の結果には非対称性が見られる. 最短距離法を用いていることにより, 結合距離 $D(d_{(IJ)K}, d_{K(IJ)})$ と $\max(d_{(IJ)K}, d_{K(IJ)})$ の間の差, あるいは, 結合距離 $D(d_{(IJ)K}, d_{K(IJ)})$ と $\min(d_{(IJ)K}, d_{K(IJ)})$ の間の差が大きく, かつ, 最後まで非対称性が残っている. ここでは　紙面の都合で省いているが, 最長距離法の場合も同様である.

この例では手法によって分類結果も各分類時の非対称性も大きく異なっている. これをどう解釈すべきかは, 種々の議論があろうし, ここでは触れないが, 強調しておきたいのは, 非対称性を無視することは大きな問題を含む場合があること, 及び, 手法の選択が重要であるということである.

表 2: 統計関連雑誌間の引用の関係

	被引用雑誌名								
雑誌名	A	B	C	D	E	F	G	H	Total
A : AnnSt	1623	42	275	47	340	179	28	57	2591
B : Biocs	155	770	419	37	348	163	85	66	2043
C : Bioka	466	141	714	33	320	284	68	81	2107
D : ComSt	1025	237	730	425	813	276	94	418	4054
E : JASA	739	264	498	68	1072	325	104	117	3187
F : JRSSB	182	60	221	17	142	188	43	27	880
G : JRSSC	88	134	163	19	145	104	211	62	926
H : Tech	112	45	147	27	181	116	41	386	1055
Total	4309	1729	3167	673	3361	1635	674	1214	16843

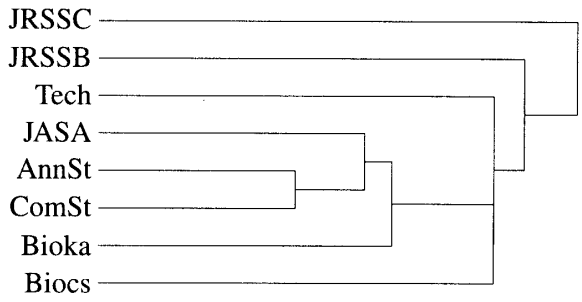

図 5: (A − a)

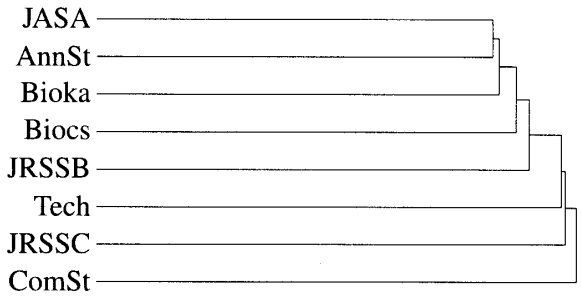

図 6: (B − a)

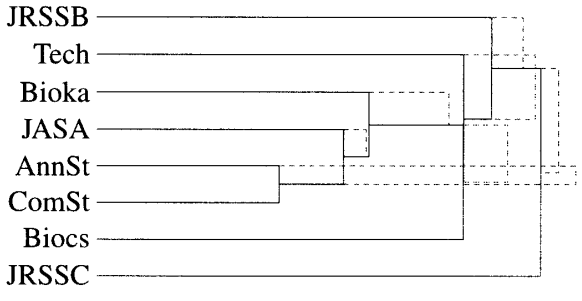

図 7: (C − c)

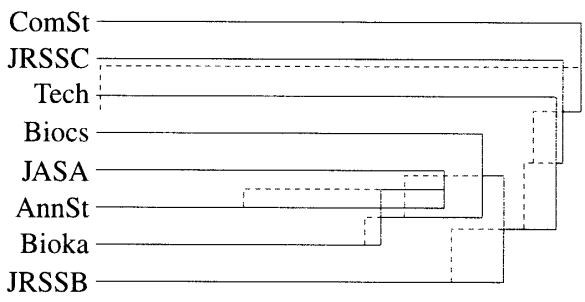

図 8: (D − c)

6. おわりに

本論文では, Lance & Williams (1967) の更新式を非対称な場合に拡張した拡張更新式を提案した. この更新式を用いることにより, 非対称凝縮型階層的クラスター分析法を対称な場合と同様, 統一的に扱うことが可能となった. また, 新たな手法の提案も容易にできるようになった. 加えて, 拡張デンドログラムにより, 結合と非対称性を同時に表示することが可能になり, 既存の表示法より解釈が容易になると考えられる.

非対称クラスター分析法の性質, 例えば, 拡張更新式のパラメータや関数と分類結果との関係などについては十分に分かっているわけではなく, 今後の検討が必要であると考えている.

謝辞

本論文の審査員の先生方には, 丁寧な査読のうえに, 不備な点のご指摘と有益なご意見を頂戴いただきました. ここに記して謝意を表します.

参考文献

Arabie, P., Schleutermann, S., Daws, J. & Hubert, L. (1988). Marketing Applications of Sequencing and Partitioning of Nonsymmetric and/or Two-mode Matrices. In W. Gaul & M. Schader (eds.), *Data Analysis, Decision Support and Expert Knowledge Representation in Marketing*, Springer Verlag, 215–224.

Brossier, G. (1982). Classification Hiérarchique à Partir de Matrices Carrées Non-symétriques. *Statistiques et Analyse des Donnees*, **7**, 22–40.

DeSarbo, W. S. & De Soete, G. (1984). On the Use of Hierarchical Clustering for the Analysis of Nonsymmetric Proximities. *Journal of Consumer Research*, **11**, 601–610.

DeSarbo, W. S., Manrai, A. K. & Burke, R. R. (1990). A Nonspatial Methodology for the Analysis of Two-way Proximity Data Incorporating the Distance-density Hypothesis, *Psychometrika*, **55**, 229–253.

Hubert, L. (1973). Min and Max Hierarchical Clustering Using Asymmetric Similarity Measures, *Psychometrica*, **38**, 63–72.

Lance, G. N. & Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems, *The Computer Journal*, **9**, 373–380.

Okada, A. & Iwamoto, T. (1996). University Enrollment flow Among the Japanese Prefectures: A Comparison Before and After the Joint First Stage Achievement Test by asymmetric Cluster Analysis, *Behaviormetrika*, **23**, 169–185.

Ozawa, K. (1983). CLASSIC: A Hierarchical Clustering Algorithm Based on Asymmetric Similarities, *Pattern Recognition*, **16**, 201–211.

Sato, M. & Sato, Y. (1995). Extended Fuzzy Clustering Models for Asymmetric Similarity, *Fuzzy Logic and Soft Computing*, World Scientific, 228–237.

Stigler, S. M. (1994). Citation Patterns in the Journals of Statistics and Probabirity, *Statistical Science*, **9**, 94–108.

岡田彬訓・岩本健良 (1995). 非対称クラスター分析法による大学進学における都道府県間の関連の分析, 理論と方法, **10**, 1–13.

藤沢秀雄 (1980). 非対称測度と等質性係数を用いたクラスタ分析, 行動計量学, **7**, 12–21.

FORMULATION OF ASYMMETRIC AGGLOMERATIVE HIERARCHICAL CLUSTERING AND GRAPHICAL REPRESENTATION OF ITS RESULT

Hiroshi Yadohisa *

* Department of Mathematics and Computer Science, Kagoshima University, Kagoshima 890-0065, Japan

Hierarchical clustering algorithms are generally based upon a (dis)similarity that is assumed to be symmetric between object pairs. However, the (dis) similarity used in actual analysis is asymmetric. Therefore, to analyze the asymmetric (dis) similarity data the researcher must perform a somehow symmetrization of his original proximity values in the beginning. On the other hand, the idea that the asymmetry has elemental meaning and the researcher must analyze the data given by using algorithm depending on the asymmetry was suggested. Hubert (1973) proposed "min and max clustering" for the asymmetric similarity. He symmetrized the data matrix in the beginning and analyze using the "min and max clustering algorithm". Fujiwara (1980) extend the Hubert's (1973) algorithm. He suggested the researcher should not perform symmetrization and should analyze the original asymmetric data matrix. Algorithms proposed in these two papers were extended the single linkage algorithm and the complete linkage algorithm to the asymmetric clustering algorithm. Okada and Iwamoto (1996) proposed the weighted average algorithm for asymmetric (dis) similarity. In those papers, they defined algorithms by deciding two steps, (i) selects the objects to be combined and (ii) updates the (dis) similarity between the objects, and not proposed uniformly.

In this paper, we define an extended updating formula to handle a profusion of asymmetric hierarchical clustering algorithms uniformly in the same manner as the symmetric one by Lance and Williams (1967). Extended dendrogram for representation of the result of analysis for asymmetric data is also proposed.

Key words: Asymmetric clustering, Asymmetric similarity, Dendrogram