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

著:宿久 洋

瀧田　孔明

富山県立大学　電子・情報工学科

July 9, 2021

はじめに

背景

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. 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.

目的

通常で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こと.

クラスター分析法の定義

2 つのプロセス

1. 結合するクラスター (対象) を選択する
2. 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

提案手法

非対称クラスター分析法の中でも凝縮型階層的なもの (AAHCA) に着目し, 今までの手法を, 対称クラスター分析法における Lance&Williams 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.

藤原の手法 (1)

対称化する場合

対称化する場合は, 前述のプロセス 1 として, 与えられた非対称類似度行列 $\mathbf{S} = [s_ij]$ 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 $\mathbf{S}^* = [s^*_{ij}]$ に変換後,$s^*_{pq} = max_{i<j}(s^*_{ij})$ を満たす対象 p,q を含むクラスター C_I,C_J を選択し, プロセス 2 として,C_I,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.

藤原の手法 (2)

対称化しない場合

対称化しない場合は, プロセス 1 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,q を含むクラスター C_I,C_J を選択し, プロセス 2 として,C_I,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.

岡田と岩本の手法

プロセス 1 は最初に対称化しない場合と同様で, プロセス 2 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.

拡張更新式

更新距離の定義

クラスター C_I,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.

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

α, β, γ:解析前に決定されている変数
f,g: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

各クラスター間距離の関係

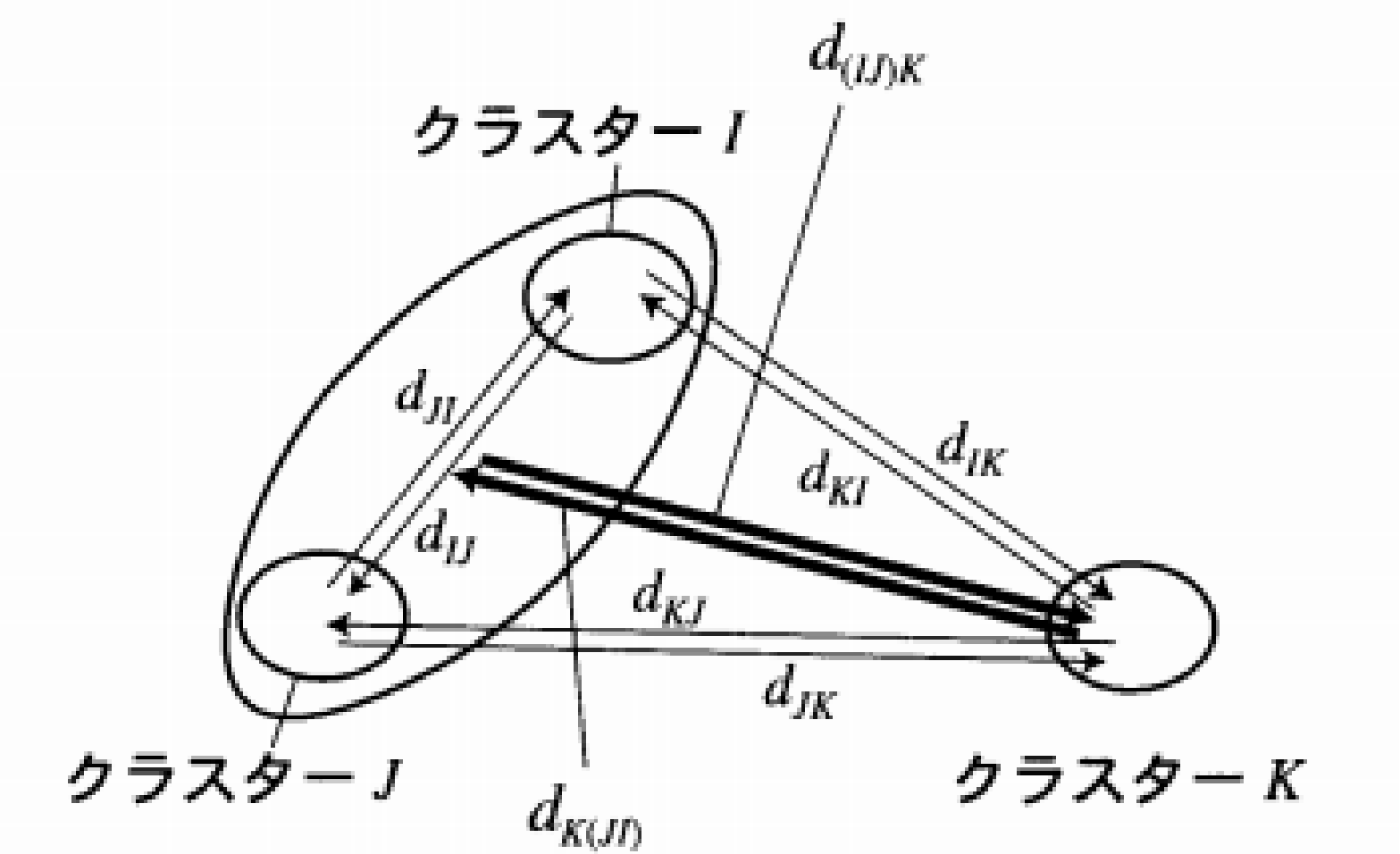

図 1: 非対称なクラスター間非類似度

拡張更新式を用いた AAHCA

非対称非類似度行列 $\mathbf{X} = (d_{ST})$ が与えられているとする.
プロセス (i):以下を満たすクラスター C_I,C_J を選択し, クラスター C_{IJ} を作る.

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

D: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
C_S, C_T, d_{ST}, d_{TS} 任意のクラスター及びそれらの間の非類似度
C_{IJ}, C_K:その段階で結合する特定のクラスター
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.

各手法の拡張更新式を用いた表現

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

岡田と岩本のデンドログラム

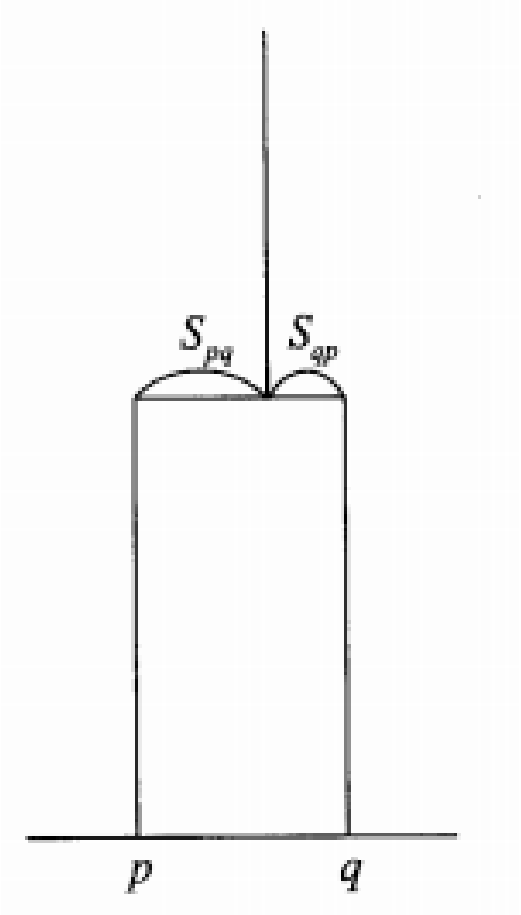

図 2: 岡田, 岩本 (1995)

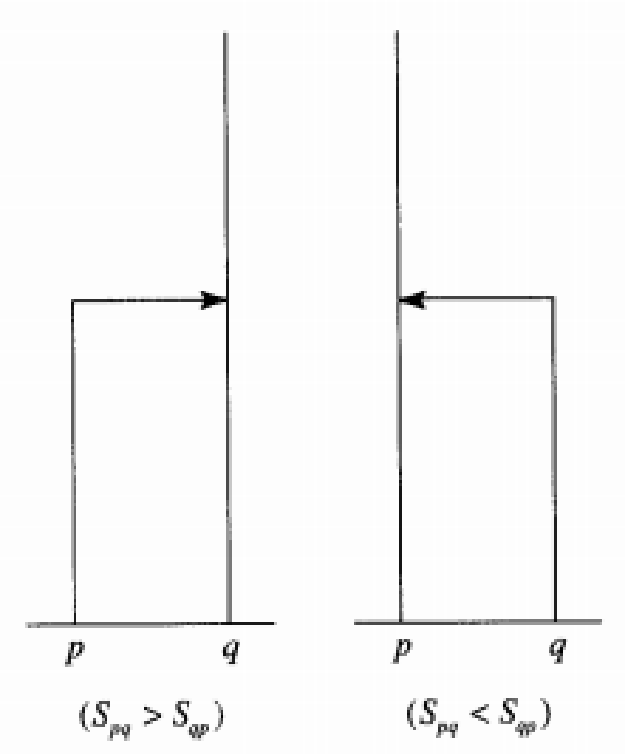

図 3: Okada and Iwamoto(1996)

拡張デンドログラム

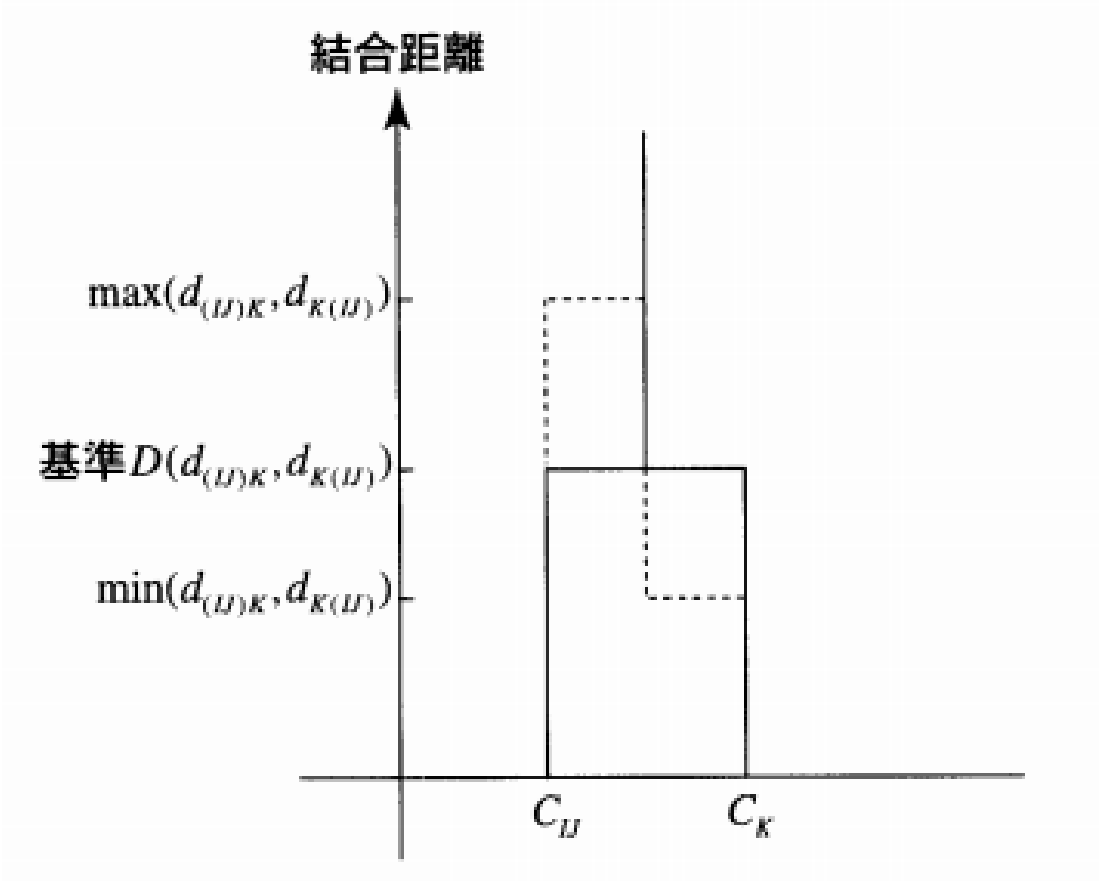

図 4: 拡張デンドログラム

拡張デンドログラムの作成 (1)

下記の表は,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.

表 2: 統計関連雑誌間の引用の関係

雑誌名	被引用雑誌名								
	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

拡張デンドログラムの作成 (2)

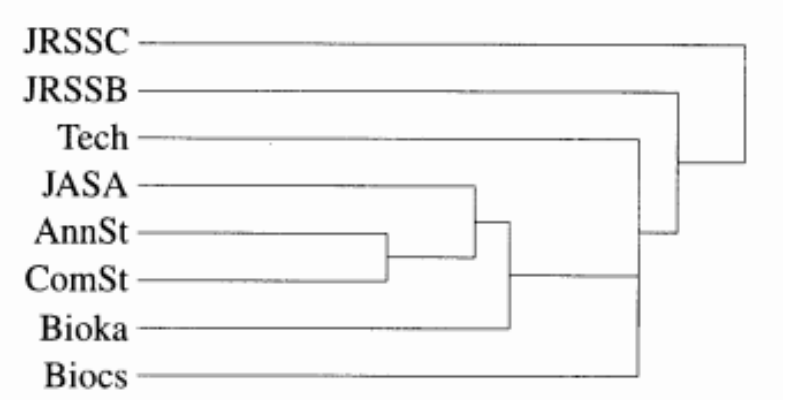

図 5: (A-a)

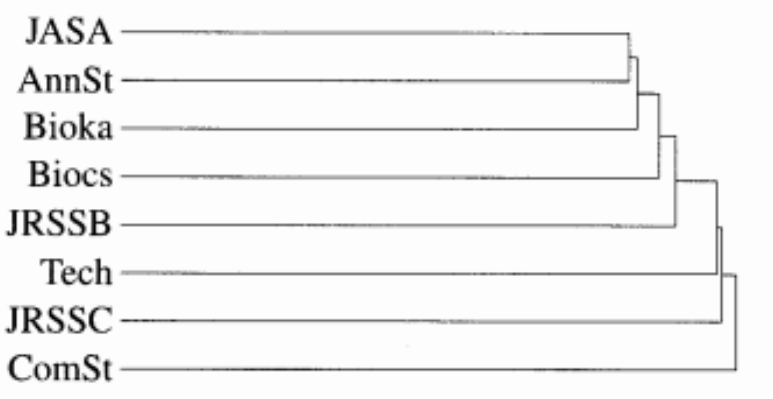

図 6: (B-a)

拡張デンドログラムの作成 (3)

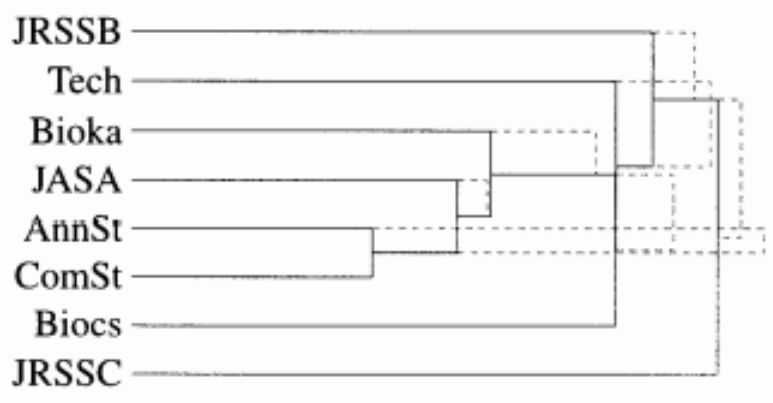

図 7: (C-c)

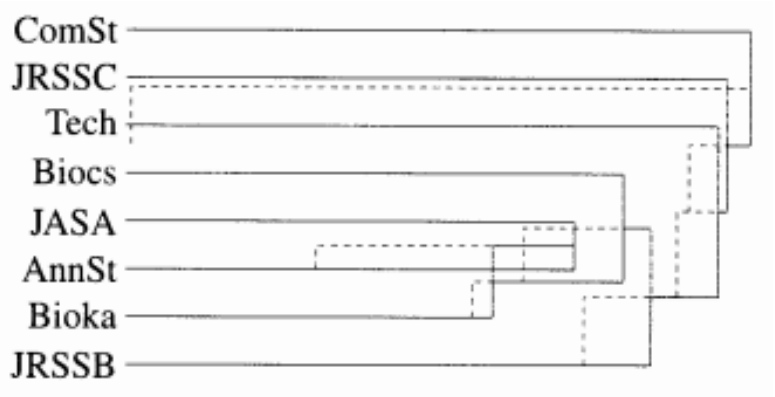

図 8: (D-c)

まとめ

可能になったこと

1. 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.
2. 新たな手法の提案も容易にできるようになった.
3. 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.

今後の課題

1. 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.