2005－CG－119（8）
2005／5／20

グラフパターンマッチングと可視化を用いた階層的クラスタリングの安定性の解析

南雲 拓　斎藤 隆文　宮村(中村) 浩子
東京農工大学 大学院 生物システム応用科学教育部

本報告では，クラスター分析の一手法である階層的クラスタリングの結果の安定性について考察する．安定性を測る手法はこれまでにもいくつか提案されているが，いずれも統計的な手法を用いている．これに対し提案手法では，幾何学的に安定性を測ることを目的とする．具体的には，階層的クラスタリングの結果として得られるデンドログラムを単純な根付き二分木とみなして，データを一つ追加しその構造の変化する領域を検出し，その領域の面積をクラスターの安定性とする．クラスターに対して安定性を定量的に表すことで，より信頼性の高い結果の表示が期待できる．

Stability Analysis of Hierarchical Clustering by Graph Pattern Matching and Visualization

Taku NAGUMO, Takafumi SAITO, Hiroko Nakamura MIYAMURA
Graduate School of Bio-Applications and Systems Engineering,
Tokyo University of Agriculture and Technology

In this report, the stability of hierarchical clustering result is considered. Conventional methods for stability analysis for hierarchical clustering uses statistical analysis. On the other hand, the proposed method aims to measure the stability with geometric analysis. A dendrogram obtained as a result of a hierarchical clustering is considered to be a binary tree. By detecting the area where an additional data causes structure change, the area is assumed to be stability of the cluster. It can expect a new expression of clustering result by quantitative stability.

1. 緒言

クラスター分析法は，複数の相関があるデータをその類似性に基づいて外的基準なしに一意に分類するための手法である．これまでにさまざまな手法が提案されており,生物学や植物学,社会科学などの分野で利用されている [1]．近年急速に発展しつつあるバイオインフォマティクス分野などで多く用いられる．

クラスター分析法は，純粋に数学的な手法であり，その性質から，データのわずかな違いによって得られる結果が大きく異なることがある．この性質は正当な推論の脆弱性となるほか，曖昧な理論の科学的裏付けとして使われる場合もあり，信頼性の意味でもクラスター分析法の安定性は重要である．

本報告では，クラスター分析法の中でも比較的よい結果が得られるとされている階層的クラスタリングを対象として，その安定性の定義を提案し，それに対する考察を行う．

以下，第 2 節で階層的クラスタリングとその安定性の関連研究について述べ，第 3 節で提案手法について述べる．第 4 節では計算機実験を行い，その結果を示す．第 5 節で結果に対する考察を行い，最後に第 6 節でまとめと今後の方針について述べる．

2. 階層的クラスタリングと関連研究

本節では，一般的な階層的クラスタリングについて解説し，その後安定性における関連研究について述べ，その問題点を指摘する．

2.1 階層的クラスタリング

n 個のデータがあるとき，距離などの類似度を用いて最も近い 2 個のデータ(あるいはクラスター)を結合する操作を $n-1$ 回繰り返すことによって，クラスターのデンドログラム(樹形図：図 1)を作成する分析法を階層的クラスタリングという．

デンドログラムの枝の長さは，データ間，クラスター間の距離を表している．階層的クラスタリングでは，あらかじめ分割クラスター数を定めなくても適当な距離で切断することによって任意の数のクラスターを得ることができる．また，用いる距離によって異なるクラスター分析法として扱うことができる．さらにデンドログラムの概形からクラスター構造，大まかなデータ間の関係などを知ることもできる．

2.2 関連研究

階層的クラスタリングの安定性に関する研究として，文献[2]では全体の階層構造について考察している．しかし経験的に，実世界においてデンドログラムが安定である場合は少ない．そこで，部分的な安定性を測る研究として，複数の階層的クラスタリングの結果間の相関測度を利用する方法がある．この相関測度を用いる方法は古くからクラスター分析法の安定性を測るために研究されてきた[3]．最近よく用いられる相関測度として，E.B.Fowlkes らによって定義された測度がある[4]．以下，この測度について詳しく述べる．階層的クラスタリングされたデータ集合 $X=\{x_1,x_2,...,x_n\}\quad (x_i\in R^d)$ を考える．ラベル L を X の k 個の部分集合 $S_1,S_2,...S_k$ のどれかを表すとする．この別の表現として行列 C で以下のように表す．

$$C_{ij}=\begin{Bmatrix}1 & x_i \,\&\, x_j (belong\quad same\quad cluster)\\ 0 & otherwise\end{Bmatrix}.$$

ラベル L_1，L_2 に対してそれぞれ行列表現 $C^{(1)},C^{(2)}$ ができ，次のように内積を定義する．

$$\langle L_1,L_2\rangle=\langle C^{(1)},C^{(2)}\rangle=\sum_{i,j}C_{i,j}{}^{(1)}C_{i,j}{}^{(2)}$$

内積 $\langle L_1,L_2\rangle$ は，コーシー・シュワルツの定理 $\langle L_1,L_2\rangle\le\sqrt{\langle L_1,L_1\rangle\langle L_2,L_2\rangle}$ を満たすので，正規化することができ，2 つのラベル間の相関測度は以下のように表すことができる．

$$cor(L_1,L_2)=\frac{\langle L_1,L_2\rangle}{\sqrt{\langle L_1,L_1\rangle\langle L_2,L_2\rangle}}$$

この相関測度を実際に用いた例として，Ben-Hur らの手法[5]があげられる．この手法は，元のデータ集合 W から，データ数が 50%以上の部分集合 W_1,W_2 $(|W_1|=|W_2|)$ を作成し，それぞれについて階層的クラスタリングを行う．このとき，共通部分 $W_1\cap W_2$ に含まれるデータに注目する．デンドログラムを 2～$|W_1|$-1 個のクラスターに分割することを考えて，それぞれの分割について共通部分のデータが W_1 と W_2 の間で所属しているクラスターが変化しているか否かを類似度として数値化する．この操作を繰り返し類似度をヒストグラムに表す．このヒストグラムの分布から最適な分割数を探すことで，安定性の高いクラスター分析結果を得ることができる．

この手法に限らず，相関測度を用いた手法において安定性とは，部分集合は元の集合と近い結果を示すという推測に基づいている．つまりこの推測部分を保証するために，異なるデータに対して繰り返し同じ処理をほどこして統計的に取り扱わなければならないという欠点がある．

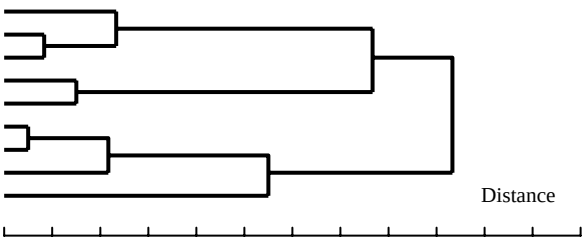


図 1 デンドログラムの例

3. 提案手法

本節では，前節で述べた統計的基準を用いずに安定性を測る手法を提案する．提案手法の特徴として，(1)デンドログラムから距離情報を破

棄し，安定性の基準としてその階層構造のみに着目する，(2)その上でデータを一つ追加し，階層構造の変化を観察する，があげられる．次項以降，これらについて詳しく述べる．

3.2 デンドログラムと安定性

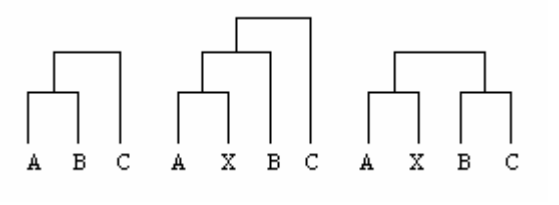
ここでは，クラスター分析が不安定な場合についてその要因を検討する．階層構造の変化が起きる要因としては，一部のデータの変化によるデータ間あるいはクラスター間の距離関係の逆転が考えられる．この距離関係の逆転は，デンドログラム上に表されている距離だけではなく，他の要因によっても引き起こされることがある．つまり，デンドログラムから読み取れる距離は安定性に関して一定の指標とはなるものの，絶対的な基準を与えない．そこで本手法では，距離情報を破棄して階層構造のみに着目して安定性を考える．

デンドログラムから距離情報を破棄し，その接続関係だけに着目すると二分木とみなすことができ，端的に階層構造を示すグラフとなる．

3.3 データの追加

前項で，階層構造が変化する要因として一部のデータの変化をあげた．クラスター分析法では大量のデータを扱うため，全データを動かして構造の変化を見ることは困難である．本手法では，新たにデータを 1 つ追加して，追加したデータを動かして階層構造の変化を観察する．

本手法の流れは次のようになる．データを 1 つ追加し，追加されたデータを含む全データに対して階層的クラスタリングを行う．追加データを動かした後，再度全データに対して階層的クラスタリングを行い，得られたグラフと 1 つ前に得たグラフを比較し，同型でない場合を検出する．これを繰り返すことで，階層構造が変化する境界線を求める．



(a)　　(b)　　(c)

図 2　階層構造の変化

追加したデータを削除することで，追加したデータのクラスタリングへの影響を調べることができる．クラスタリング結果から追加データをその上位階層への枝とともに削除する．これによって得られたグラフとデータを追加する前の結果のグラフを比較し，同型でない場合には，追加したデータそのものに関係しない階層構造の変化を検出できる．このとき，階層構造の変化を起こす追加データ位置の領域が小さいほど，クラスタリング結果は安定であると言える．例えば，図 2(a)のような構造があるとき 1 点追加することを考える．組合せは多々あるが，代表として(b)と(c)をあげる．この場合，(b)では追加点である X を除いて階層構造が変化していない．逆に(c)は X を除いた後の階層構造が変化している．

2 つのグラフが同型であるかどうかの判別をグラフパターンマッチングと呼び，グラフが二分木である場合には以下の方法で判別できる．グラフ $T=(V,E)$ を任意の二分木，根を r とし，r から最も遠いデータまでの道の長さを T の高さと呼ぶ．また，r から各頂点 $v \in V$ までの道の長さを v の高さと呼ぶ．これらが，

（ｉ） T_1, T_2 の全頂点数が同一，高さが等しい

（ii） 各頂端点 $v \in V$ が同一の深さを持つ

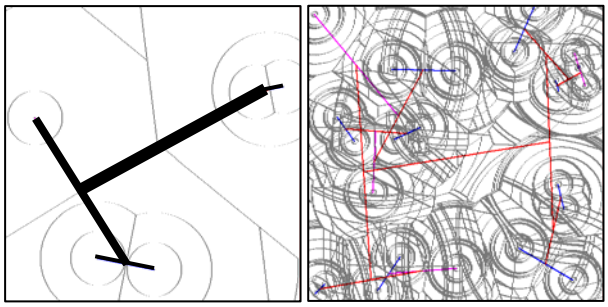
を満たせば，2 つのグラフは同型である．提案手法においては，比較するグラフの頂点数は常に同一であるため，（ｉ）については高さの比較のみで済む．

4. 対話的なツールによる実験

階層構造の変化を対話的に調査するために，GUI を備えた階層的クラスタリングを行うツールを作成した．追加点を動かしつつ逐次クラスタリングを行い，その変化を読み取ることができる．

ここでは簡単のため，ここでは二次元平面に配置された点を考える．データ間の非類似度はユークリッド距離として，またクラスター間非類似度は重心法を用いて階層的クラスタリングを行う．

構造が変化する境界線を抽出するために，追加データを動かしつつグラフパターンマッチングを行い，グラフが同型でない点を検出する．



(a) データ数:5　　(b) データ数:25

図 3　構造変化の検出結果

データはラスタ順に動かす．データ数が 5，25 それぞれの場合の結果を以下に示す(図 3).

ここで，直線あるいは曲線は階層構造の変化境界を表している．また，太い線分の端点が各データである．階層構造が上位になるにつれてより太い線分として表現している．データが 1 つのクラスター同士の結合がもっとも細く，複数のクラスター同士の結合をもっとも太く，1 つのデータを持つクラスターと複数のデータを持つクラスターの結合は中間で表現している．

図 3(a)からは，それぞれのデータ，クラスターの勢力範囲を容易に読み取ることができる．また図 3(b)では，データ数が多く，変化境界が多すぎて一見して対応がわかりづらい. そこで，次節ではデータ数が少ない場合について考察する．

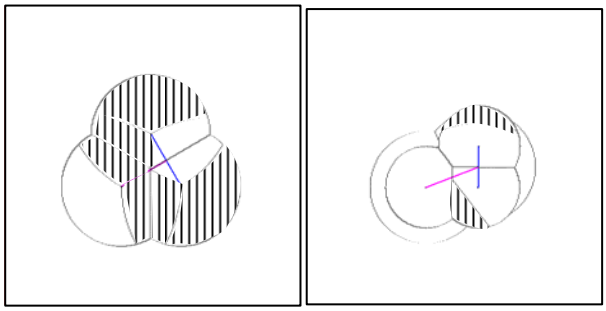
5.　データ数が 3 の場合の理論的考察

本節ではデータ数が 3 である場合について, 理論的考察を行う．

5.1　数値実験

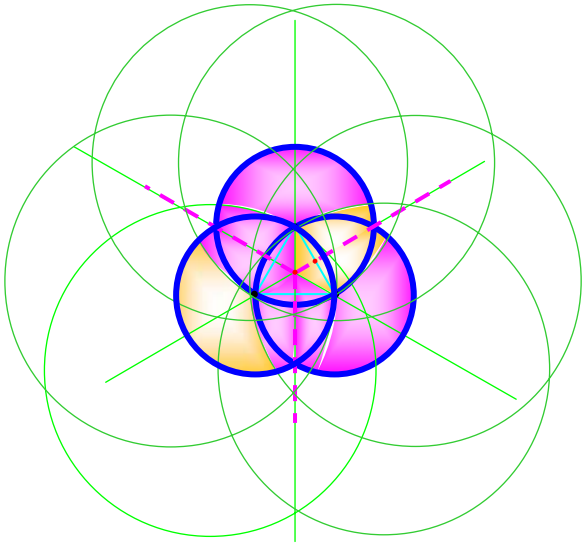
データ数を 3 とし，(a)正三角形に近い形，(b)辺の比が 3:4:5 の三角形，の 2 例について検討する．追加データの削除を行い，追加データに関係しない階層構造が変化する部分(3.3 項)について数値実験を行った結果を図 4 に示す．図 4 の網掛け領域は追加データに関係せず階層構造が変化している領域である．

同じ例について，後述(5.2, 5.3 項)の方法を用いて作図したものが，図 5 である．図 5 の濃い

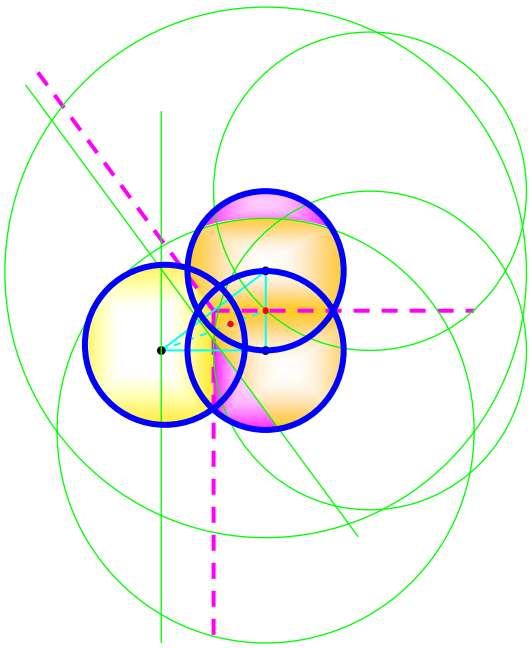


(a)正三角形に近い形　　(b)3:4:5 の三角形

図 4　3 点の場合(実験)



(a)正三角形に近い形



(b)3:4:5 の三角形

図 5　3 点の場合(理論)

網掛け領域は,階層構造が変化した部分である．これらから各データ間の距離の差が小さいと変化が起こりやすく，逆に偏った配置であると変化は起こりにくいことがわかる．これは，距離の差が小さい場合にはデータ間の距離の逆転が生じやすいためである．安定性の基準として，黄色く示された部分の面積比が利用できると考えられる．面積は，境界を作る要素である円，直線の方程式から算出することができる．

以下，これらの円および直線の意味と方程式について述べる．

5.2 考察すべき対象領域

階層構造の変化が起きる境界線は，円と直線からなる．図 5 に描かれている境界線はそれぞれ以下のような意味を持つ．

(青)の太線：追加したデータが最寄の点と先の結合する境界線
(紫)の太破線：ボロノイ線
(緑)の細線：追加したデータが最寄の点と先に統合したとき,他の点との距離の大小関係が変化する境界線

これら境界線を式で表すために，3 点を O，A，B とおき，座標を図 6 のようにおく．

これらのうち 2 点 A，B が互いに結合する以前に，追加データが 3 点 A，B，O のうちいずれか 1 点と結合するための条件は，追加データが以下の円内部に存在することである．

a. A を中心とする円：

$$(x-x_a)^2+(y-y_a)^2=c^2 \quad (1)$$

b. B を中心とする円：

$$(x-x_b)^2+(y-y_b)^2=c^2 \quad (2)$$

c. O を中心とする円：

$$x^2+y^2=c^2 \quad (3)$$

これらの内部の領域を R とする．

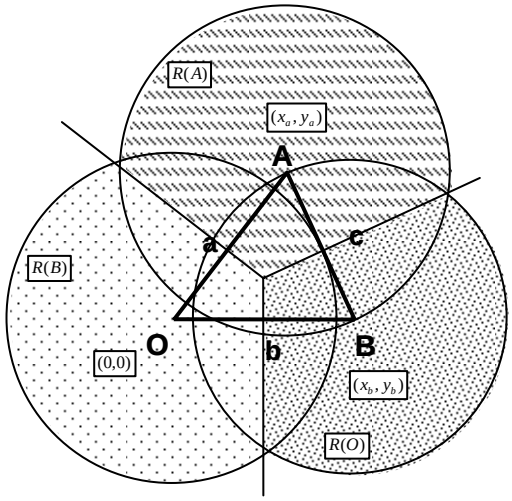
さらに，3 本のボロノイ線により領域 R は 3 つの領域に分割される．

d. OA 間のボロノイ線：

$$x_a x+2y_a y=a^2 \quad (4)$$

e. OB 間のボロノイ線：

$$x_b x+2y_b y=b^2 \quad (5)$$



ただし，OA＞OB＞AB

図 6　座標の定義と領域 R

f. AB 間のボロノイ線：

$$(x_a-x_b)x+2(y_a-y_b)y=a^2-b^2 \quad (6)$$

分割された領域をそれぞれ $R(A)$，$R(B)$，$R(O)$ とする．これらの領域は，追加データがどの点と結合するかを示しており，階層構造の変化が起こりうる領域である．また追加データが最初のステップで A, B, O と結合したとき，そのクラスターの重心をそれぞれ A’，B’，O’ とおく．

5.3 階層構造が変化する条件

前項で定めた領域 $R(A)$，$R(B)$，$R(O)$ の内部に追加データが入るとき一定の条件を満たすと階層構造が変化する．以下それぞれの領域についてその条件について述べる．なお，境界線は次のルールで命名する．

X_{YZ} ：X が Y，Z どちらに近いかの境界線

(i) $R(A)$

$R(A)$ の領域で問題となるのは，A’，B，O の距離関係である．$R(A)$ 内部で A’，B，O の距離関係の境界線は，

円：$B_{A'O}$

$$\{x-(2x_b-x_a)\}^2+\{y-(2y_b-y_a)\}^2=4a^2 \quad (7)$$

円：$O_{A'B}$

$$(x+x_a)^2+(y+y_a)^2=4a^2 \quad (8)$$

直線：A'_{BO}

$$x_b x+y_b y=b^2-x_a x_b-y_a y_b \quad (9)$$

であり，データ追加に関係しない階層構造の変化が起こる条件は，

・円 $B_{A'O}$ の外側

または

・直線 A'_{BO} で二分される領域のうちの O 側となる．

(ii) $R(B)$

以下同様に，

円：$A_{B'O}$

$$\{x-(2x_a-x_b)\}^2+\{y-(2y_a-y_b)\}^2 = 4b^2 \quad (10)$$

円：$O_{B'A}$

$$(x+x_b)^2+(y+y_b)^2=4b^2 \quad (11)$$

直線：B'_{AO}

$$x_a x+y_a y=a^2-x_a x_b-y_a y_b \quad (12)$$

この場合の変化が起こる条件は，

・円 $A_{B'O}$ の外側

または

・直線 B'_{AO} で二分される領域のうちの O 側となる．

(iii) $R(O)$

同じように，

円：$A_{O'B}$

$$(x-2x_a)^2+(y-2y_a)^2=4c^2 \quad (13)$$

円：$B_{O'A}$

$$(x-2x_b)^2+(y-2y_b)^2=4c^2 \quad (14)$$

直線：O'_{AB}

$$(x_a-x_b)x+(y_a-y_b)y=a^2-b^2 \quad (15)$$

変化が起こる条件として，

・円 $A_{O'B}$ の内側

または

・円 $B_{O'A}$ の内側

となることがわかる．

6. 結言

本報告では，任意のデータを一つ追加することで，安定性を測る手法を提案した．

また，データ数が少ない場合について追加データに関わらず階層構造が変化している領域を実験的に確認した．また，境界線を作る要素を理論的に求め，安定性の条件を示した．今後は，データ数が多い場合についても検討する．このとき，デンドログラム全体に対してデータを一つ追加する方式と，低い階層部分について適用し，徐々に上位階層にあげていく方式の 2 パターンを考えている．前者はデータ数に応じて境界を構成する要素も増えるため，実現は困難であることが予測される．しかし後者では，データ数が大きいグラフは分割することによってデータ数が少ない場合のグラフの集合と考えられる．このことを利用し，分割された低い階層部分から提案手法を再帰的に適用することで比較的容易に結果を得ることができると考えられる．

また，定量的な安定性をグラフ構造上に可視化する新たな階層的クラスタリング結果の表示方法についても検討する．

参考文献

[1] A.K. JAIN, M.N. MURTY, and P.J. FLYNN,“Data Clustering: A Review,” *ACM Computing Surveys*, Vol.31,no.3,pp.264-323, 1999.

[2] S.P. Smith and R. Dubes, “Stability of a hierarchical clustering,”*Pattern Recognition 12*,pp.177-187,1980.

[3] V.V. Ranghavan and M.Y.L.IP, “Techniques for measuring the stability of clustering:a comparative study,” *ACM SIGIR 1982*, pp.209-237, 1982.

[4] E.B. Fowlkes and C.L. Mallows, “A method for comparing two hierarchical clusterings,” *Journal of the American Statistical Association 1983*, pp.553-584, 1983.

[5] A. Ben-Hur et al,“A stability based method for discovering structure in clustered data,” *Pacific Symposium on Biocomputing*, pp.6-17, 2002.