2006－MPS－58（9）
2006／3／16

仮想要素追加法による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，階層的クラスタリング結果の安定性を解析するための数理モデルを提案する．このモデルでは，従来手法のような統計的処理を用いずに，仮想要素の追加によって幾何学的に安定性を測ることができる．提案法を 2 次元ユークリッド空間でのクラスタリングに適用し，階層安定度を樹形図上に可視化することで，その有効性を検証する．

Stability Analysis of Hierarchical Clustering by Adding a Temporary Element

TAKU NAGUMO，TAKAFUMI SAITO，HIROKO NAKAMURA MIYAMURA
Graduate School of Bio-Applications & Systems Engineering
Tokyo University of Agriculture and Technology

In this report, a mathematical model is proposed for analyzing the stability of hierarchical clustering results. In this model, the stability is measured geometrically by adding a temporary element, without using a statistical analysis. The proposed method is applied to clustering examples in 2-dimensional Euclidean space, and the effectiveness is verified by mapping the hierarchical stability onto the dendrogram.

1. 緒言

クラスタ分析法は，複数の相関を持つ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，生物学や社会科学などの分野で利用されている [1]．特に近年は，バイオインフォマティクス分野において不可欠な技術となっている．

クラスタ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．そのため，クラスタ分析を仮説の科学的裏付けなどに使う場合には，クラスタリング分析結果の安定性を考慮に入れることが重要である．しかし現実には，得られた結果の安定性に対して考察が行われることは少ない．その理由として，クラスタ分析手法の普及に比べて，その安定性に関する研究がまだ十分とはいえず，特に安定性を手軽に求める手法が開拓されていないことがあげられる．

本報告では，クラスタの適切な分割数が未知のときに用いられる階層的クラスタリングを対象として，その安定性を解析するための数理モデルを提案する．安定性の指標として，従来手法が元のデータ集合やその部分集合におけるクラスタの類似性を用いているのに対し，提案手法では，元のデータ集合に仮想的な要素を追加した場合の階層構造の変化の有無に着目する．それにより，統計的手法を用いることなく個々の階層ごとの安定度の算出が可能となる．さらに，階層安定度を樹形図上に可視化する手法についても提案する．

2. 階層的クラスタリングの安定性

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，安定性の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 個の要素データをもつデータ集合に対して，最も近い 2 個の要素（あるいはクラスタ）を結合する操作を $n-1$ 回繰り返すことによって，クラスタの樹形図を作成する分析法を，階層的クラスタリングという．樹形図の枝の長さは，要素，あるいはクラスタ間の距離を表している．階層的クラスタリングでは，あらかじめクラスタ分割数を定めなくても，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を得ることができる．また，樹形図の概形からクラスタ構造，大まかな要素間の関係などを知ることもできる．

2.2　安定性の関連研究

階層的クラスタリングの安定性に関する研究としては，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代表的である[2]．たとえば，Corneil らは，Rand の分類間類似測度[3]を安定性に用いている[4]．また，Yu はグラフ理論的に安定性を測る手法を提案している[5]．近年よく用いられる類似測度として，Fowlkes らによって定義された測度がある[6]．この測度を実際に用いた例として，Ben-Hur らの手法[7]があげられる．この手法では，元のデータ集合の部分集合をランダムに 2 つ作成し，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，2 つの部分集合の共通部分に含まれる要素に注目する．樹形図をクラスタに分割することを考え，それぞれの分割について共通部分の要素の所属しているクラスタが変化しているか否かを類似度として数値化し，統計的な処理を行って安定なクラスタ分割を得る．

これら既存手法は分類間の類似測度によるため，統計的に用いなければならないという欠点がある．

3. 仮想要素追加法による安定性モデル

本節では，統計的基準を用いずに安定性を測る手法を提案する．また，提案法を 2 次元ユークリッド空間に適用した例を示す．

3.1　仮想要素追加法による安定性のモデル化

本手法では，元のデータ集合に対し，要素を新たに 1 個追加して階層的クラスタリングを行い，その位置による階層構造の変化を検出する．追加要素を加えてクラスタリングし，そのうえで樹形図から追加要素を削除することで，追加要素のクラスタリングへの影響を調べることができる. 追加要素の削除は，追加要素をその結合対象に同化させることで実現する．得られたクラスタ構造と，要素追加前のクラスタ構造を比較し，同一でない場合には，本質的な階層構造の変化とみなす．いま，図 1(a)のような 3 要素からなるクラスタ構造があるとき，要素 P を追加してクラスタリングを行うことを考える. このとき，たとえば(b)のような構造になった場合は，追加要素である P を除くと，階層構造は(c)に示すように(a)と変化していない．これに対して，(d)のような構造になった場合は，P を除いた後のクラスタ構造は(e)のように変化しており，本質的な階層構造変化であることがわかる．

要素の追加によって，上記のような本質的な階層構造変化が起こるか否かは，追加要素の値に依存する．このとき，階層構造変化を引き起こすような

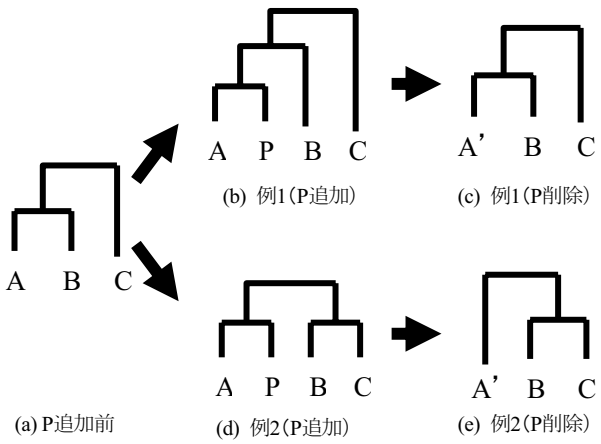


図 1　仮想要素 P の追加削除による階層構造の変化

追加要素値の範囲が大きいほど，そのクラスタ構造は不安定であると考えることができる．

3.2　階層安定度の定義

前項で述べた追加要素値の範囲によるクラスタ構造の安定さを定式化し, 階層安定度として定義する．ここでは，追加要素 P が A，B，C いずれかの要素と先に結合する場合だけを対象として考え，そのときの P のとりうる値の範囲を領域 R(n)とする．たとえば，図 1(b)，(d)となる場合は，いずれも P が A と結合するので, そのときの P の値は R(n)に含まれる．領域 R(n)は，本質的な階層構造変化が起こる領域 R(u)と，起こらない領域 R(s)に分けられる．このとき，R(n)に占める R(s)の領域の大きさの割合，すなわち R(s)／R(n) を，A，B，C の 3 要素からなるクラスタの階層安定度と定義する．

なお，A，B，C は，その一部もしくは全部がクラスタであっても，その代表値を用いることで，同様に階層安定度を定義できる．ただし，簡単のため，それぞれのクラスタは十分安定であり仮想要素の追加によって崩壊しないという仮定を設ける．

3.3　2 次元ユークリッド空間における適用例

安定度を実際の空間に対して適用した結果を示す．ここでは簡単のため 2 次元空間を対象とする．要素間の距離尺度はユークリッド距離とし，クラスタ間距離は重心法とする．

3 要素 A，B，C の配置として，各要素間距離が $|AB|:|AC|=1:\sqrt{2}$ の場合と $|AB|\fallingdotseq|AC|\fallingdotseq|BC|$ の場合について考える．それぞれの仮想要素追加法による安定度を，近似的に計算する．R(n) 領域内部の各画素について，本質的な階層構造変化が起きるか否かを判定することで，R(s)，R(u) 領域の画素を数え上げる．R(u)領域に着色した結果を図 2 に示す．(a)の安定度は 0.88 であり，(b)では 0.34 となる．

ここで仮に，(b)の場合で|AB|=|AC|=|BC|として図 3 のように R(n)を分割することを考えると，R(At)，R(Bt)，R(Ct)の面積はそれぞれ等しく，また R(Al)，R(Ar)，R(Bl)，R(Br)，R(Cl)，R(Cr)の面積もそれぞれ等しい．ここから階層安定度の値域は次のようになる．

$$\frac{1}{3} \leq \text{階層安定度} \leq 1$$

先に述べた図 2(b)の場合の安定度 0.34 は，最も不安定なときの理論値 1/3 に近い値となっている．

さらに詳しく安定度について見るために，先に結合する 2 要素 A，B を固定し，3 個目の要素を A，B それぞれを中心とする半径|AB|の外部で動かし，安定度の分布を調べる．この結果を図 4 に示す．ここで，白領域は走査範囲外であることを示している．3 要素間の距離がほぼ等しくなる 2 円の交点近辺で，安定度は特に低くなり，距離差が大きくなるにつれて安定度が高くなっていることが読み取れる．

4. 実験・考察

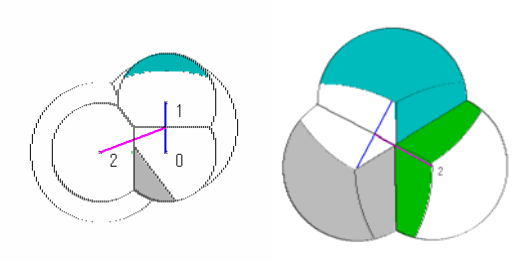
本節ではランダムに作成したデータ群を例として，仮想要素追加法による安定度によって有効に表現される情報，および表現されない情報とその解決策について述べる．また，従来手法との比較についても考察する．

4.1 提案手法の有効性

提案法では，仮想要素の追加による階層構造の変化を不安定としているが，これはすでに形成されたクラスタに対してその内部の要素や他クラスタの要素が微小な変動をした場合を仮想要素に見立てることを意味している．

この様子を確かめるため，直観的に理解しやすい 2 次元データを例として，実際のデータと樹形図から提案法の有効性を確かめる．要素数を 20 としランダムに作成した例について，図 5 に座標，図 6 に階層的クラスタリング後に安定度を求め，それを輝度値で示した樹形図を示す．ここで，安定度の輝度値へのマッピングには図 4 と同じものを用いる．また計算対象となる 3 クラスタは，結合順序にしたがい先に結合しているクラスタの代表値と，後に結合するクラスタの 2 つの子の代表値とする．安定度は，その値を持つノードから，その子である後に結合するクラスタまでの範囲の矩形を安定度に対応する輝度値で塗りつぶして表現する．

これらの要素のうち，図 5 右側にある{(4, 16), (6, (5, 15))}からなるクラスタに着目する(図 7)．(4, 16)と(6, (5, 15))間には十分な距離があるように見える．ここ



(a)|AB|:|AC|=1:$\sqrt{2}$ (0.88)　(b)|AB|≒|AC|≒|BC| (0.34)

図 2　構造変化領域 (括弧内は安定度)

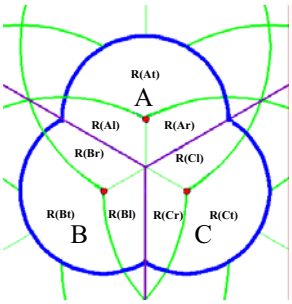


図 3　対象領域

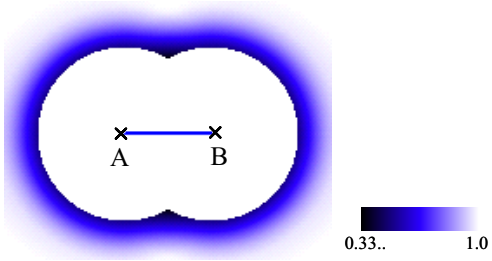


図 4　安定度分布

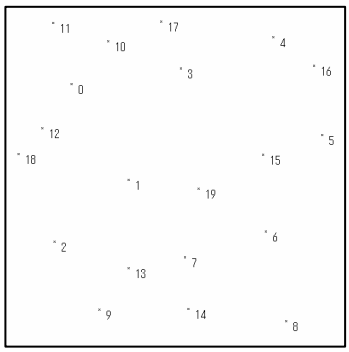


図 5　2 次元データ集合（要素数：20）

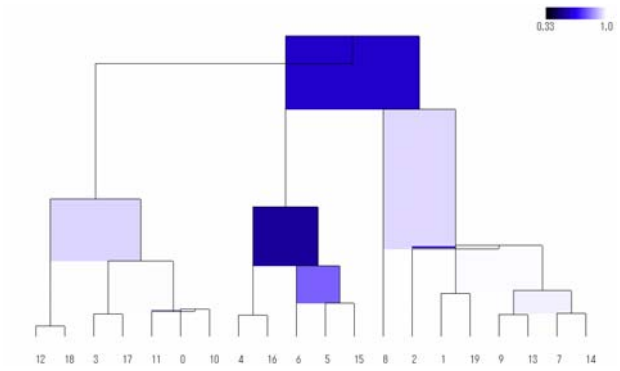
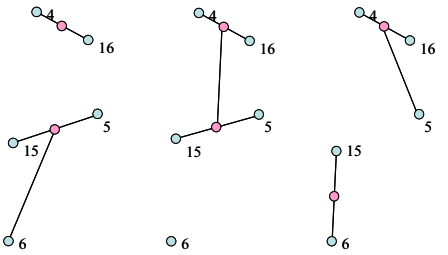


図 6　階層安定度を可視化した樹形図（要素数：20）



(a) 変化前　(b) 上へ変化後　(c) 下へ変化後

図 7　要素 15 が微小に変化する場合

で要素 15 が上方向に微小移動すると，クラスタ(5, 15)の代表値と要素 6 との距離よりも(4, 16)との距離が短くなり先に結合し，クラスタ{(4, 16), (5, 15)}となってしまう．また下方向へ微小移動すると要素 6 と要素 15 の距離が要素 5 とクラスタ(4, 16)との距離よりも小さくなりクラスタ構造が変化する．

本手法は，すでにクラスタリングされた結果である樹形図に対して階層安定度を計算する．つまり安定度計算対象を各ノードに対してその下位 3 ノードに定めているため，計算対象から漏れる要素・クラスタ間の逆転可能性は示すことができない．このことへの対応として，すでに結合しているノードだけでなく，結合する可能性のある近傍ノードすべてに対して安定度を計算することが考えられる．しかし，計算量や可視化手法などの問題は残る．

4.2　既存手法との比較

本手法はデータ集合の部分集合を用いる手法などと比べて統計的に扱う必要がないため，計算量は著しく減少する．階層安定度は，現時点では画素の数え上げで算出しているが，対象 3 クラスタのクラスタ間距離だけから計算できるため，表引きと補間によりさらに高速な近似計算が可能と考えられる．

本手法の優位点として，データ数が少ない場合にも適用できることがあげられる．部分集合を用いる手法では，データ数が一定以上ない場合には十分な信頼性を得られないのに対し，本手法はわずか 3 個の要素からなるクラスタに対しても，安定度を計算できる．

また本手法をクラスタ分割数決定の指標として利用することも考えられる．従来，分割数を決定するためには結合距離を用いているが，これに各階層における安定度を加味して考慮することで，より適したクラスタ分割が得られると考えられる．

5.　結言

本報告では，仮想要素を追加することで，階層的クラスタリングの安定性を幾何学的に解析する新しい数理モデルを提案した．本手法では，ランダムサンプリングによる統計的手法を用いることなく，各階層での安定度を算出できる．また，階層安定度を樹形図上に可視化することで，従来の樹形図では失われていた情報を提示できた．

今後の課題として，現実のアプリケーションに適用し，有効性を検証することがあげられる．今回は，2 次元ユークリッド空間で重心法を用いた場合に限定して，階層安定性を算出した．しかし，種々のアプリケーションデータに適用するには，多次元空間で，種々の距離尺度，種々のクラスタ間距離を用いた場合に対しても，対応が必要である．

さらに，階層安定度の高速計算，階層を越えた安定性の解析を行うための計算法と可視化法，クラスタ分割数の決定法，より直観的な可視化手法の開拓などにも，順次取り組みたい．

謝辞

本研究の一部は，科学研究費補助金（萌芽 17650024）の援助を受けている．

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] 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.

[3] W. M. Rand, “Objective criteria for the evaluation of clustering, ” *Journal of American Statistical Association*, Vol. 66, No. 336, pp. 846-850, 1971.

[4] D. G. Corneil and M. E. Woodward, “A comparison and evaluation of graph theoretical clustering techniques,” *INFOR, Canadian Journal of Operational Research and Information Processing*, Vol. 16, No. 1, pp. 74-89, 1978.

[5] C. T. Yu, “The Stability of two common matching functions in classification with respect to a proposed measure,” *Journal of the American society for Information Science*, Vol. 27, No. 4, pp. 248-255, 1976.

[6] E. B. Fowlkes, and C. L. Mallows, “A method for comparing two hierarchical clusterings,” *Journal of the American Statistical Association*, Vol. 78, No. 78, pp. 553-584, 1983.

[7] A. Ben-Hur, A. Elisseeff, and I. Guyon, “A stability based method for discovering structure in clustered data,” *Pacific Symposium on Biocomputing*, Vol. 7, pp. 6-17, 2002.