2007－MPS－63 (2)
2007／3／3

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と可視化

渡部 秀文[1] 南雲 拓[2] 一宮 和正[3] 斎藤 隆文[1] 宮村(中村) 浩子[1]

[1] 東京農工大学 大学院 生物システム応用科学府
[2] 東京農工大学 大学院 生物システム応用科学府(現 株式会社リコー)
[3] 東京農工大学 工学部 情報コミュニケーション工学科

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新しい数理モデルを提案する．また安定性とクラスタ要素の広がり度合いを可視化してクラスタの最適な分割数を補助的に求める手法について提案する．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．一方，クラスタ分割を決定するための指標として，クラスタ要素の広がり度合いについて述べ，階層安定度と要素の広がり度合いを樹形図上に可視化する手法についても提案する．また，提案手法と従来手法にサンプルデータを適用し，提案手法の有効性及び問題点を比較検証する．

Stability Analysis and Visualization of Hierarchical Clustering by Adding a Temporary Element

HIDEFUMI WATANABE[1], TAKU NAGUMO[2], KAZUMASA ICHIMIYA[3], TAKAFUMI SAITO[1], HIROKO NAKAMURA MIYAMURA[1]

[1] Graduate School of Bio-Applications & Systems Engineering Tokyo University of Agriculture and Technology
[2] Graduate School of Bio-Applications & Systems Engineering Tokyo University of Agriculture and Technology (Presently with Ricoh Company, Ltd.)
[3] Department of Computer, Information and Communication Sciences, Tokyo University of Agriculture and Technology

We propose a new mathematical model for analyzing the stability of hierarchical clustering results. In this paper, a method for deciding the most suitable number of clusters with visualization of stability and density of cluster elements is also proposed. In this model, the stability is measured geometrically by adding a temporary element, without using a statistical analysis. In this method, we focus on the change of hierarchical structures when an element is added. On the other hand, the density of clusters elements as an indicator for deciding the dividing of the cluster is presented. Moreover, the method to visualize stability and density of the elements of the clusters is proposed. We demonstrate the effectiveness and problems of the proposed method by applying it to the sample data.

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，データのわずかな違いにより得られる結果が大きく異なることがある．そのため，クラスタ分析を科学的裏付けなどに使う場合には，分析結果の安定性を考慮に入れることが重要である．

本報告では，クラスタの適切な個数が未知のときによく用いられる階層的クラスタリングを対象とし，その安定性を解析するための新しい数理モデルを提案する．提案手法では，データ集合に仮想的な要素を追加した場合の階層構造変化の有無に着目する．さらに，クラスタ分割を判断する指標として，クラスタ要素の広がり度合いの可視化手法について述べる．一般的に，階層的クラスタリングでは結合対象のクラスタの下層の要素については，要素の広がり度合いが反映されない．そこで，要素の広がり度合いを階層安定度とともに樹形図上に可視化して，最適な分割数を求める手法についても提案する．さらに，従来手法との比較及び検証も行う．

2. 階層的クラスタリングの安定性

データ集合に対して，最も近い2個の要素またはクラスタを結合する操作を繰り返して，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．階層的クラスタリングでは，樹形図を適当な階層で切断することで任意の数のクラスタを得ることができる．

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]. 例えば，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相関測度として，Fowlkes らによって定義された測度がある[6]. この相関測度を実際に用いた例として，Ben-Hur らの手法が挙げられる[7]. この手法は，元

のデータ集合から，部分集合をランダムに作成し，それぞれ階層的クラスタリングを行う．このとき，共通部分に含まれる要素に注目する．樹形図を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要素の所属しているクラスタが変化しているかを類似度として数値化し，安定なクラスタ数を得る．これらの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．

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．

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 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．要素を加えてクラスタリングし，樹形図から追加要素を削除することで，追加要素の影響を調べることができる．得られたクラスタ構造と，追加前のクラスタ構造を比較し，同一でない場合は，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．図 1 (b)のような構造になった場合は，P を除くと，階層構造は(c)に示すように(a)と同一である．これに対して，図 1 (d)のような場合は，P を除いた後の構造は(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，本質的な階層構造変化が起こるかは，追加要素の値に依存する．階層構造変化を引き起こす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

追加要素値の範囲によるクラスタ構造の安定さを階層安定度として定義する．追加要素 P が A，B，C いずれかと先に結合する場合の P 値の範囲を領域 R(n)とする．例えば，図 1(b)，(d)となる場合は，いずれも P が A と結合するので，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 R(s)/R(n) を，A，B，C からなるクラスタの階層安定度と定義する．A，B，C は，その一部もしくは全部がクラスタであっても，階層安定度は同様に定義できる．ただし簡単のため，A，B，C 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

2 次元空間に適用した結果を示す．要素間距離はユークリッド距離，クラスタ間距離は重心法とする．3 要素からなるクラスタに対して，追加要素が階層構造変化を引き起こす様子を解析する．

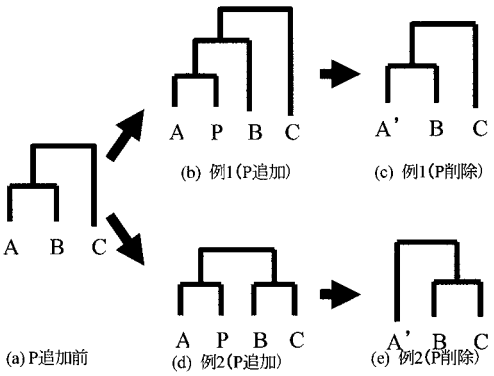
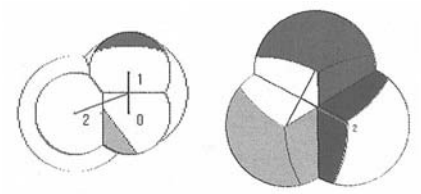


図 1 仮想要素 P の追加削除による階層構造変化



(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)

要素 A，B，C の要素間距離が|AB|:|AC|=1:$\sqrt{2}$ と|AB|≒|AC|≒|BC|の場合を考える．仮想要素追加法による安定度を近似的に計算する．R(n) を描画し，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を判定し，R(s)，R(u) の画素を数え上げる．R(u)領域に色を付けた結果を図 2 に示す．

3.4 コサイン距離の適用

現実のケースで多く用いられる，コサイン距離，群平均法を用いたクラスタリングへ適用する．コサイン距離は，正規化した要素ベクトルの内積である．群平均法は，クラスタ間の距離を，両クラスタ間の要素の全ての対で要素間距離を求め平均し求める．

3.4.1 安定度の考え方

安定度は 3.3 項同様 R(n)内の R(u)の割合で定義する．コサイン距離では，要素ベクトルを正規化するため，n 次元データであれば，正規化した要素は n 次元単位超球面上に分布する．そこで，R(n)等の大きさは，超球面上の(n-1)次元超体積として求める．

3.4.2 安定度の求め方

多次元データでは，安定度を幾何学的に求めることは困難であるため，単位超球面上で R(n)となる点をサンプリングにより求め，安定度を算出する．

(1) 単位超球面上に等密度に点を配置し，R(n)内の点であるか調べる
(2) R(n)内である場合，この点を要素に追加して階層構造変化を調べ，R(s)か R(u)かを判定する
(3)点の個数比から，安定度 R(s)/R(n)を求める

3.4.3 計算の高速化

サンプリング法では，データの次元が高くなると計算量は指数的に増大するため，高速化について述べる．安定度は，クラスタそれぞれの(1)要素数と，正規化した要素の平均を代表点としたときの，(2)代表点同士の内積の関数で表される．そこで(1)，(2)の値を一定間隔で変更したときの安定度を計算し，表にしておくことで，表引きと補間で安定度の近似値を高速に求めることができると考えられる．

4. 階層安定度及び広がり度合の可視化

本節では，提案した階層安定度と要素の広がり度合いを樹形図上に可視化する手法を提案する．

4.1 各階層の安定度の可視化

安定度を樹形図に表現するために，安定度計算対象の 3 ノードを結んで三角形を描画し，安定度に割り当てた色で塗る．図 4 に，100 個のデータ集合から得た樹形図に安定度と広がり度合いを可視化した例を示す．安定度は，色が白いほど不安定である．

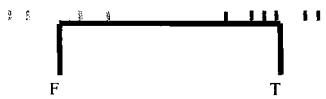
4.2 クラスタ要素の広がり度合いの可視化

階層安定度には要素の広がりが反映されていないため，クラスタを分割すべきでない場合にも高い安定性を示すことがある．クラスタを分割できるかを判別するために，次のようにクラスタ要素の広がり度合いを可視化する．

(1) 分離後の 2 個のクラスタの各代表値を求める
(2) 全ての要素を，代表値を結ぶ直線上に投影する
(3) 投影結果を樹形図上にクラスタごとに色分けして描画する

図 3(a)に，クラスタが二つに分かれる場合の要素の投影のイメージを，(b)に可視化例を示す．

(a)投影の考え方



(b) (a)の要素の樹形図への可視化イメージ
図 3 要素の広がり度合いの表現

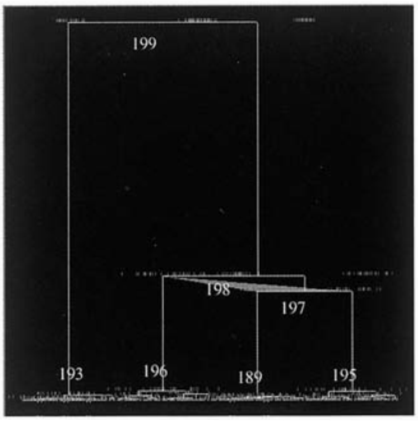


図 4 要素の広がり度合いの可視化結果 (100data)

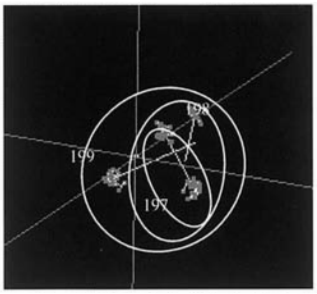


図 5 サンプルデータ集合(100data)

4.3 提案手法によるクラスタ分割数の決定

クラスタの分割の可否は次のように判断できる．
(1)赤と緑の分布が分離していないとき，分割不可
(2)クラスタを構成する 3 個のサブクラスタ間での安定度が低いとき，分割不可

さらに，クラスタ対の一方に比べて他方の要素が極端に少ない場合，特異点と考え分割しない．

例として，図 4 クラスタの分割数を求めると，分割数は 193，198 の 2 分割か，193，196，189，195 の 4 分割が適当である．図 5 に元となった 3 次元データを示す．なお，最終的な分割数は，アプリケーションやデータ特性に依存するため，ユーザによって都度判断されるべきである．

5. 既存手法との比較実験

本節では既存手法として Ben-Hur らの手法[7]を実装し，提案手法の有効性，問題点を検証する．第 5 節で使用したデータを使用し，①プログラム実行時間，②結果のわかりやすさを見る．

5.1 実験結果

表 1 に実験結果を，Ben-Hur 法の実行結果を図 6 に示す．図 6 から，安定度が高く最多の分割数 4 が結果となる．提案手法の結果は，5.4 項でも示したとおりで，分割数は 4 となる．

表 1　実験結果

	実行時間(秒)	クラスタ数(個)/正否
提案手法	150	4 / 正
Ben-Hur 法	40	4 / 正

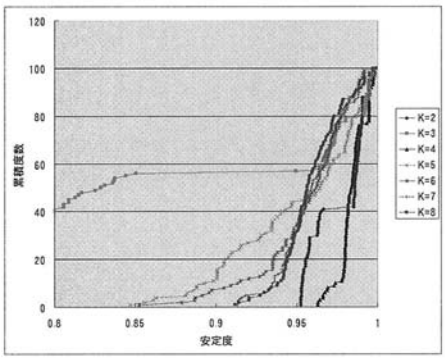


図 6　Ben-Hur 法の結果

5.2　プログラム実行時間

提案手法は，サンプリング法を用いているため時間がかる．しかし，3.4 項で述べたように，表引きと補間により，高速化の余地はある．Ben-Hur 法は，十分な試行回数が必須であり，高速化は困難である．

5.3　結果のわかりやすさ

Ben-Hur 法では，クラスタ分割数は図 6 のように読み取ることができるが樹形図との対応が取れない．そのため，個々のクラスタを得るには樹形図を分割数になる層で切るしかない．提案手法では，樹形図上に必要な情報が提示されるので，階層は固定されず，特定のクラスタを深く分割することもできる．

6.　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．また，階層安定度と要素の広がり度合いの可視化結果を用いたクラスタ分割手法を提案した．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，クラスタ要素の広がり度合いと安定度を可視化することで，よりわかりやすくクラスタを分割できる．

今後の課題として，計算時間の効率化・高速化が挙げられる．現実の例に適用し，有効性を検証することも挙げられる．また，3.2 項で除外した，クラスタ構造が破壊されるケースへの対応も行う．

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.