社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－CG－132 (12)
2008／8／23

# 階層的クラスタリングの安定性の可視化

一宮 和正† 渡部 秀文‡ 宮村(中村) 浩子‡ 古谷 雅理† 斎藤 隆文‡
†東京農工大学 大学院 工学府 情報工学専攻
‡東京農工大学 大学院 生物システム応用科学府

本研究では，階層的クラスタリングの安定性を可視化し，クラスタ分析に役立つ情報を提示する．クラスタ分析に利用される階層的クラスタリングには，ノイズや誤差など，データの僅かな違いによって結果が大きく異なることがある．そのため，階層的クラスタリングでは，結果の信頼性を示す安定性を考慮に入れる必要がある．本研究では，安定性の測定手法の一つとして，安定性を各階層で幾何学的に測定することができる仮想要素追加法を用い，部分木内での安定性の可視化と，部分木をまたがる安定性の可視化手法を提案する．

# Visualization of Stability of Hierarchical Clustering

Kazumasa Ichimiya,† Hidefumi Watanabe,‡ Hiroko Nakamura Miyamura,‡
Tadasuke Furuya,† and Takahumi Saito‡
†Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology
‡Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology

We propose a method to visualize stability of hierarchical clustering result. The hierarchical clustering used for the cluster analysis has the problem that the result often changes according to a little difference of input data like the noise and the error, etc. Therefore, we should consider cluster stability for reliable result. We use the "Adding a Temporary Element Method" for the stability determination. And, we propose the method of visualization of the hierarchical structure that uses the dendrogram and the method of visualization of the stability that uses two-dimension table.

## 1. はじめに

階層的クラスタリングを用いたクラスタ分析法は，複数の相関を持つデータをその類似性に基づいて一意に分類する手法である．しかし，階層的クラスタリングには，ノイズや誤差などのデータの僅かな違いによって，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に使う場合には，分析結果の安定性を考慮に入れる必要がある．

現在，階層的クラスタリング結果の安定性を測定する手法は多く提案されているが，その研究目的は，最適な分類数を得ることにとどまり，クラスタ構造を十分に分析できる情報が得られない問題がある．そこで，仮想要素を追加したときのクラスタ構造の変化に着目して安定性を求める仮想要素追加法[1]が提案された．仮想要素追加法では，安定性を各クラスタ，または各階層で求めることができる．しかし，ここで求まる安定性は数値として表されるため階層構造のどの領域がどの程度の安定性であるかを直感的に認識するためには，安定性の可視化が必要不可欠である．

階層的クラスタリング結果の可視化には一般的に距離付樹形図が使われている．しかし，距離付樹形図により提示できる情報は，階層構造とクラスタの結合距離(類似度)のみである．さらに，階層的クラスタリンの安定性を可視化した研究は少なく，安定性を使ったクラスタ分析のためには，新たな可視化手法の提案が必要である．本研究では，仮想要素追加法を使った階層的クラスタリングの可視化手法を提案し，クラスタ分析に役立つ情報を提示することを目的とする．

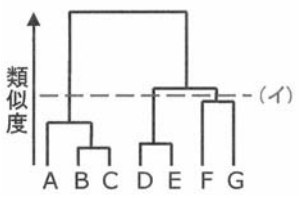


図 1: 距離付樹形図

## 2. 関連研究

本章では，一般的に階層的クラスタリング結果の可視化に使われる距離付樹形図と，階層的クラスタリングの安定性を可視化した Ben-Hur らの手法 [2]について述べる．また，バイオインフォマティクス分野において使われる階層的クラスタリングの可視化手法について述べる．

### 2.1. 距離付樹形図

階層的クラスタリングでは，ただ 1 つの階層構造が得られる．その階層構造を可視化する一般的な手法として距離付樹形図がある．距離付樹形図では，階層的クラスタリングの階層構造を表現するだけでなく，クラスタ，または要素間の結合距離(類似度)を階層の深さにマッピングして表現する(図 1)．

クラスタ分析において，距離付樹形図の結合距離の長い部分で樹形図を分け，クラスタの最適数を得る方法もあるが，単に大きなクラスタの場合にも結合距離が長くなることから，この方法は信頼できないといわれている．そのため，クラスタの最適数を求める場合には，結合距離だけでなく，結果の安定性を考慮する必要がある．しかし，距離付樹形図からは安定性の情報が得られず，距離付樹形図のみでのクラスタ分析は，誤った結論を導く恐れがある．

### 2.2. Ben-Hur らの手法

Ben-Hur らは，E.B.Fowlkes らの類似性測度 [3]を用いて階層的クラスタリングの安定性を測定し，クラスタの最適数を決定する手法を提案している．ここで Ben-Hur らは階層的クラスタリングの安定性をヒストグラムにより可視化している．

Ben-Hur らは，安定性の測定結果から分割を得るときに 2.1.節の距離付樹形図を使う．例えば，階層的クラスタリングの結果，図 1 のような階層構造が得られ，ヒストグラムから分割数 3 で安定性が高いとわかった場合，図 1(イ)の部分で樹形図を分け，(A,B,C)，(D,E)，(F,G)という分割を得る．

しかし，この手法は，部分集合は元の集合と似た結果を示すという推測に基づいているため，統計的に扱わなくてはならず，直感的な理解が難しい．また，ヒストグラムにより安定性を可視化しているが，このヒストグラムは距離付樹形図とは独立した形で表現され，そこから得られる情報はクラスタの最適数のみである．そのため，この Ben-Hur らの手法では，どの要素がどのクラスタに含まれるのかといった詳細な情報を得るためには，さらなる分析が必要である．

### 2.3. バイオインフォマティクス分野での可視化

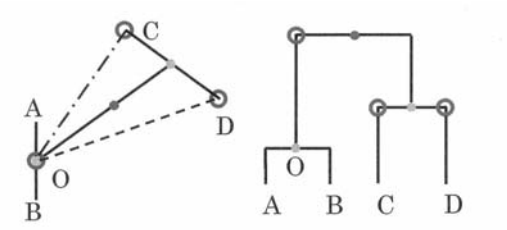
階層的クラスタリングを使ったクラスタ分析は，遺伝子分析などバイオインフォマティクス分野において多く利用されている．クラスタリング結果の可視化には，一般的に距離付樹形図が使われるが，そこから得られる情報は少なく，分析が困難である．そこで，樹形図と同時に，クラスタ分析に必要な情報を合わせて提示する手法が必要となる．J. Seo らは，樹形図のリーフに遺伝子データを合わせて可視化することで，樹形図と元データとの対応を容易にした [4]．さらに，データを 2 次元座標にマッピングした結果や，違うクラスタリングアルゴリズムでの結果を比較する手法なども提案されている．

しかし，クラスタ分析において得に重要である安定性を可視化する研究は行なわれていない．さらに，各クラスタ，または階層における安定性を可視化する場合，樹形図のノード間の関係を可視化する必要があり，既存の可視化手法では困難である．

## 3. 提案手法

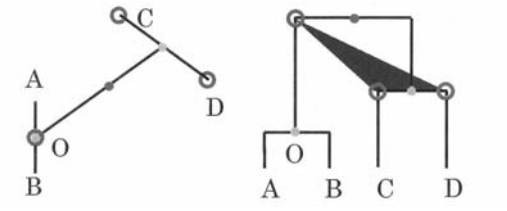
本章では，階層的クラスタリング結果の可視化手法として，樹形図の作成，樹形図上への可視化 [1]，2 次元表による可視化の手法について述べる．また，安定性は仮想要素追加法 [1](付録)により測定し，図 2 の輝度にマッピングする．

図 2: 安定性のマッピング




(a)特徴空間の座標　　(b)樹形図

図 3: 樹形図のノード配置




(a)特徴空間の座標　(b)樹形図上へ可視化

図 4: 樹形図上への安定性の可視化

### 3.1. 樹形図の作成

階層的クラスタリング結果は，距離付樹形図で可視化するが，そのノード配置を提案する．

階層的クラスタリング結果である階層構造を樹形図により可視化するにあたって，特徴量空間におけるクラスタ，または要素間の距離(類似度)を考える．クラスタの代表点，または要素が図 3(a)のように分布する場合，A，B が先に結合し，次に C，D が結合する．このクラスタ構造を樹形図に可視化するとき，A+B のクラスタ代表点 O と C，D との距離(類似度)を考える．図 3(a)では，|OC|<|OD|であることがわかる．そこで，樹形図でも C を O に近い側に配置することで，距離(類似度)の関係を樹形図上に可視化する(図 3(b))．

このノード配置では，全要素の距離(類似度)の関係は考えず，あくまで樹形図上の局所的な 3 点の関係のみでノードの位置を考える．そのため，樹形図全体で考えると，必ずしも正しい距離の関係になっているとはいえない．しかし，3 点の距離関係を樹形図で表現することで，その 3 点でのクラスタ構造の安定性が低い場合，どのような構造変化が起こりやすいかがわかる．例えば，図 3(b)のような樹形図があり，O，C，D の間で安定性が低い場合，D より C の方が先に O と結合するような構造変化が起こりやすいことがわかる．

### 3.2.樹形図上への安定性の可視化

著者らは，安定性を 3 要素，またはクラスタから計算する仮想要素追加法を使い，樹形図上に安定性を可視化する手法を提案している [1]．

樹形図上に可視化するには，クラスタリングの結合順序によりノードを選択し安定性を計算，可視化する．データが図 4(a)のように分布する場合，A と B が先にクラスタを形成するため，図 4(b)のように A+B のクラスタ O と要素 C，D の間で安定性を計算し可視化する．可視化には，どの 3 ノードについての安定性であるかがわかるように，その 3 点を頂点とした三角形で可視化する(図 4(b))．図 4(b)の場合，O，C，D の間で安定性が低いことがわかり，C と D のいずれかが，先に O と結合して階層構造の変化が起こる可能性があることがわかる．

さらに，3.1.節で提案したノード配置により，C が O と先に結合する変化の可能性の方が高いことがわかる．

### 3.3. 2 次元表による安定性の可視化

樹形図に表現されるクラスタ構造だけでなく，より多くのクラスタ，または要素間での安定性を可視化する．そこで，より多くの情報を提示するため，樹形図とは別の可視化手法を考える．ただし，クラスタ分析の手間を軽減するため，樹形図との対応の行ないやすさも考える．

まず，ある階層で存在するクラスタ間の安定性を考える．そこで，階層を指定し，その階層でクラスタを形成する 2 要素，またはクラスタとその他の要素，またはクラスタの間でそれぞれの安定性を計算し可視化する．列を可視化する階層に存在する要素，またはクラスタ，行をクラスタを形成する 2 要素，またはクラスタの組とし，表の形で安定性を可視化する．さらに，その並び順と位置を樹形図に合わせ，樹形図との対応を容易にする．

例えば，図 5 では，可視化する階層に存在する要素，またはクラスタが 9,3,4,11,12 で，この階層でクラスタを形成する組は(3,4)と(11,12)であり，それぞれの安定性を計算し，可視化している．このとき，(3,4)と 3 など，自分自身を含むものとの安定性は考えず，ただ太線の枠を付けて表示する．

また，行，列，のノードの位置を完全に樹形図に一致させた表を複数の階層で作成し，乗算合成で重ねた，重ね表も提案する(図 6)．これにより，複数の階層の可視化結果を 1 つの可視化結果にまとめることができる．

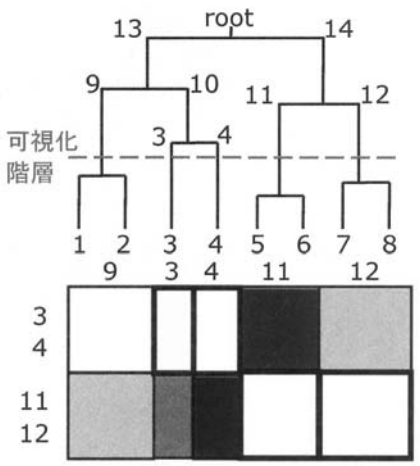



図 5: 2 次元表による安定性の可視化

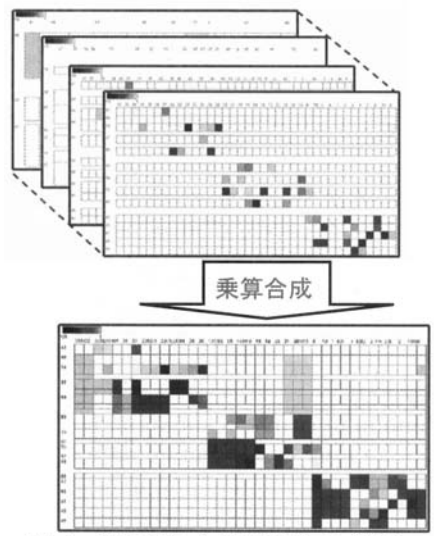



図 6: 重ね表による安定性の可視化

## 4. 実験

提案手法を，3 種類の人工データと，実データとして気象データを適用し，考察する．

### 4.1. データ

人工データは，2 次元座標空間上にプロットして人工的に発生させた，一様分布(図 7(a))，2 つのガウスの混合分布(図 7(b))，3 つのガウスの混合分布(図 7(c))に近い分布となるデータを使用した．実データは，気象庁報道発表資料 2006 年 95 地点の年間平均気温と年間降水量のデータ(図 8)を使用した．

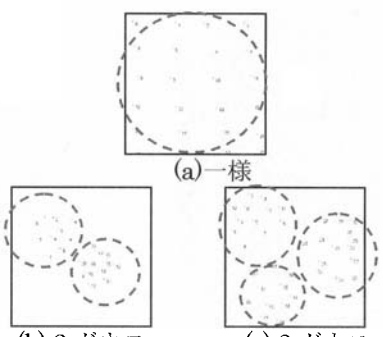



図 7: 入力データの分布図


図 8: 気象データ(年間平均気温と降水量)

### 4.2. 考察

実験により得られた可視化結果(図 9,10,11)について考察を行なう．

#### 4.2.1. クラスタ分析

樹形図上への可視化結果(図 9,11(a))から，階層的クラスタリング結果として求められた階層構造には，多くの不安定な構造があることがわかる．クラスタ分析において，安定性の低い部分を深く分析することで，新しい発見が得られる可能性も考えられる．

また，2 次元表による可視化結果(図 10,11(b))からは，樹形図に表れるクラスタ構造以外のクラスタ間の安定性が可視化され，どのクラスタ間で影響を及ぼしあい，構造が不安定になるかという情報が得られる．図 11(b)の(イ)の部分では，樹形図上では全く異なるクラスタに属していると考えられるクラスタ(図 11(a)(ロ))と要素(図 11(a)(ハ))の間で安定性が低いことがわかる．これは，それらのクラスタ間の距離(類似度)が実際は非常に近く，互いに近い特徴を持っている可能性があるからで，元データである図 8 からも(ロ)，(ハ)の位置が近いことがわかる．このように，これまで一般的であった樹形図による階層的クラスタリング結果の可視化結果ではわからなかった情報を提示することができる．

#### 4.2.2. クラスタ最適数の推定

人工データでの実験結果の2次元表による可視化結果(図 10)から，安定性の分布と，予め設定したクラスタ数(ガウスの数)が一致していることがわかる．特に，重ね表による可視化結果(図 10(d)(e)(f))からは，その特徴を明確に見ることができ，クラスタ分析において重要なクラスタの最適数を推定することが可能であるといえる．また，図 10(f)の(二)の部分の安定性が僅かに低くなっていることから，その部分のクラスタ，または要素は，所属するクラスタ以外へも変動する可能性が考えられる．このように，クラスタの最適数を得るだけでなく，さらに多くのクラスタ構造の可能性を考えることができる．

## 5. まとめ

階層的クラスタリングの安定性を樹形図上とそれに関連した2次元表の形で可視化した．2 次元表による可視化からは，クラスタの最適数の推定が行なえた．さらに，2 つの可視化手法を組み合わせて使うことで，これまでのクラスタ分析では考えることができなかった，より深い分析を行なうことが可能となった．

今後の課題として，実データのように，より複雑なデータに対して分析に多くの手間がかかることが挙げられる．そこで，樹形図と2 次元表を統合した可視化手法の検討を含め，より直感的な可視化手法が必要であるといえる．また，クラスタ分析のために，さらに多くの情報を可視化する必要がある．

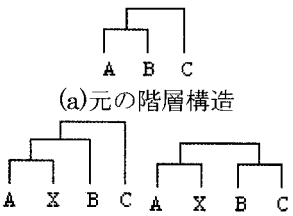
### 参考文献


1) 渡部 秀文, 南雲 拓, 一宮 和正, 斎藤 隆文, 宮村(中村) 浩子, "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," 情報処理学会論文誌:数理モデル化と応用, Vol.48, No.SIG15(TOM18), pp.176-188, 2007.
2) A. Ben-Hur, A. Elisseeff, and I. Guyon, "A Stability Based Method for Discovering Structure in Clustered Data," In *Proc. Pacific Symposium on Biocomputing 2002*, pp.6-17, 2002.
3) E. B. Fowlkes and C. L. Mallows, "A Method for Comparing two Hierarchical Clusterings," *Journal of the American Statistical Association*, Vol.78, No.383, pp.553-569, 1983.
4) J. Seo and B. Shneiderman, "Interactively Exploring Hierarchical Clustering Results," *IEEE Computer Society Press*, Vol.35, No.7, pp.80-86, 2002.


## 付録: 仮想要素追加法

仮想要素追加法では，クラスタリング対象が3点である場合を考え，ここに仮想要素を1 つ追加し，クラスタリングを行なったときの構造変化に着目する．例えば，図 12(a)のような階層構造に要素Xを追加してクラスタリングを行なった場合，図 12(b)や(c)などの結果が得られる．図 12(b)，(c)ともに追加したXがAとクラスタを形成しているが，特に図 12(c)に注目するとBがAより先にCとクラスタを形成している．このように，追加要素と直接関わらない部分の構造が変化する場合に，階層構造に変化が起きたと定義する．

次に階層構造の変化から安定性を測定する手法について述べる．図 13 では点A，B，Cをクラスタの代表点，または要素とする．このとき，まずAとBがクラスタを形成する．そこで，それぞれの点が半径|AB|の領域を持つとし，これら全てを合わせた領域をRとする．仮想要素を領域 R 内に追加した場合，いずれかの点とクラスタを形成し，階層構造の変化が起こる可能性がある．領域 R 中で，実際に階層構造に変化が起こる領域は図 13 の着色された領域となる．そこで，仮想要素の追加により階層構造が変化する可能性のある領域Rに対し，実際に変化が起こる領域の比率を計算し，この比率をこの3点による階層構造の安定性とする．この安定性は最も不安定な場合で 1/3，安定な場合で1となる．



(a)元の階層構造
(b)安定　(c)不安定
図 12: 階層構造の変化

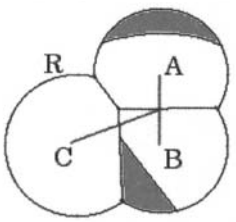
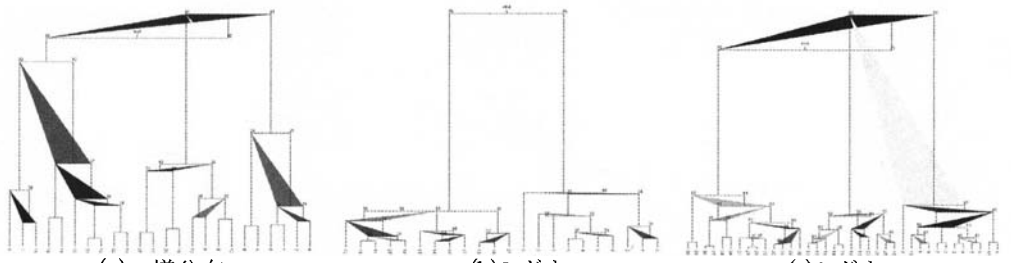
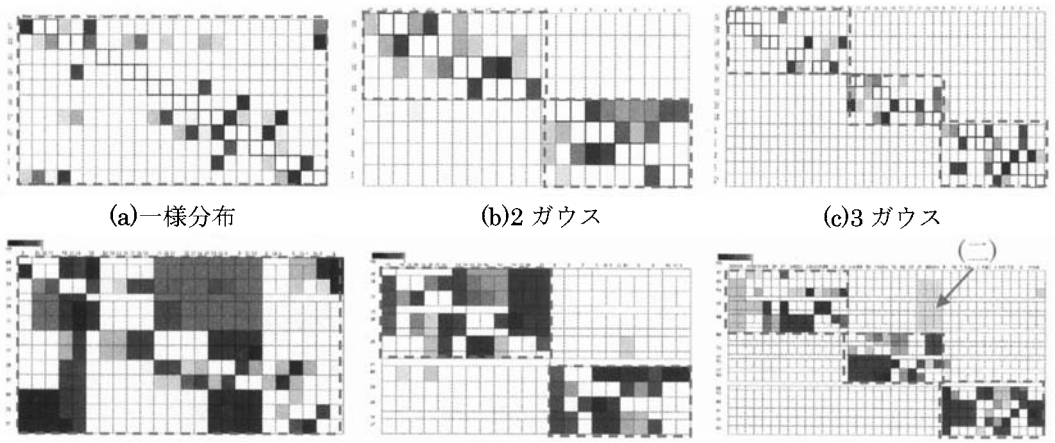


図 13: 安定度計算領域



(a)一様分布　(b)2 ガウス　(c)3 ガウス

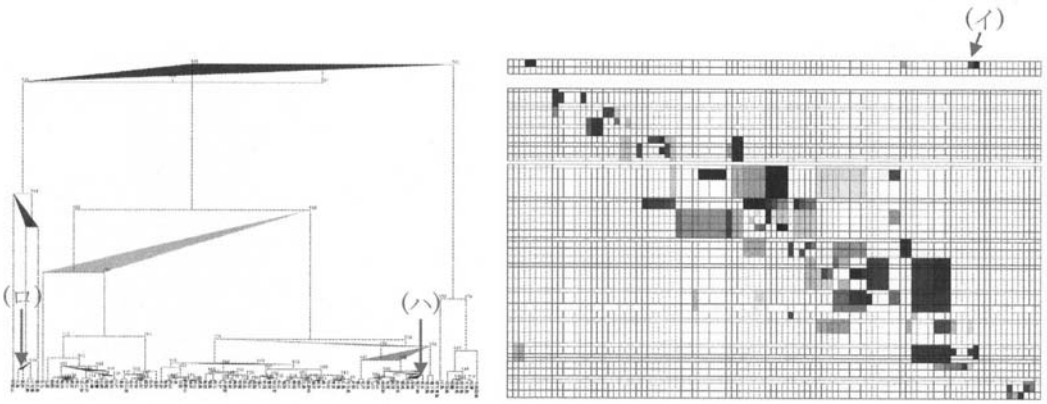
図 9: 樹形図上への安定性の可視化結果




(a)一様分布　(b)2 ガウス　(c)3 ガウス

(d)一様分布(重ね表)　(e)2 ガウス(重ね表)　(f)3 ガウス(重ね表)

図 10: 2 次元表による安定性の可視化結果




(a)樹形図上への可視化　(b)2 次元重ね表による可視化

図 11: 気象データの可視化結果