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[†] 南雲 拓[†,☆] 一宮 和正[††]
斎藤 隆文[†] 宮村（中村）浩子[†]

本論文では，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求める手法について提案する．階層的クラスタリングは，未知のデータ集合から意味のある分類を得る目的でしばしば用いられる．しかし，結果の安定性に関する研究は十分なされているとはいえず，安定性を手軽に求める手法も開拓されていない．本論文では，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手法を提案する．この手法では，要素を 1 個追加して階層的クラスタリングを行い，得られた結果の階層構造変化に着目する．追加要素の位置によって，本質的な階層構造変化が起こる場合と起こらない場合とがある．そのうち，構造変化が起こらない要素の割合を算出することで階層安定度を得る．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る．さらに，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，提案手法の有効性および問題点について比較検証する．

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

Hidefumi Watanabe,[†] Taku Nagumo,[†,☆] Kazumasa Ichimiya,[††]
Takafumi Saito[†] and Hiroko Nakamura Miyamura[†]

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. Hierarchical clustering is often used in order to obtain meaningful classification from an unknown dataset. However, the stability of the clustering results is not studied enough, and the techniques for simply calculating the stability measure have never been developed. In this paper, 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. If there is more stable region of the added element without structure change, the structure is more stable. In this context, the hierarchical stability is obtained by calculating the ratio of the stable area. 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．これまでに様々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

† 東京農工大学大学院生物システム応用科学府
Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology
†† 東京農工大学大学院工学府情報工学専攻
Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology
☆ 現在，株式会社リコー
Presently with Ricoh Company, Ltd.

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．

本論文では，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て，その安定性を解析するための新しい数理モデルならびに可視化手法を提案する．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変化の有無に着目する．それによって，ランダムサンプリングによる統計的手法を用いずに，個々の階層ごとの安定度の算出が可能となる．本論文では，提案手法をユークリッド距離・重心法およびコサイン距離・群平均法の 2 種類の距離尺度で適用する．また，クラスタを分割するもう 1 つの指標として，クラスタ要素の広がり度合いについて述べる．一般的に，階層的クラスタリングでは上層においてはクラスタの代表値を求めてそれらの距離に着目する．その際，下層に属している要素については，クラスタどうしの距離を求める際に重みとして反映されることはあるが，広がり度合いについては反映されない．そこで，クラスタ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．また，従来手法との比較および検証も行う．

本論文の構成は次のとおりである．まず 2 章で，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る．3 章では，仮想要素追加法による安定性モデルを提案し，階層安定度を定義する．また，2 次元ユークリッド空間で重心法を用いた場合を例として，具体的な安定度の計算方法ならびに適用例を示す．4 章では，クラスタの広がり度合いの定義と，その可視化手法に関して述べる．5 章では，得られた階層安定度とクラスタ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．また，例を用いてクラスタの分割決定法を提案する．6 章で従来手法との比較実験と，その考察について述べる．最後に 7 章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の安定性

本章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 個の要素データを持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，あらかじめ分割クラスタ数を定めなくても，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，大まかなデータ間の関係などを知ることもできる．

階層的クラスタリングでは，要素間の距離（非類似度とも呼ぶ）の定義と，クラスタ間の距離の定義に，それぞれ複数の方法が考えられる．これらの選択によって，異なるクラスタ分析法として扱うことができる．

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相関測度として，Fowlkes らによって定義された測度がある[6]．以下では，この測度について詳しく述べる．ある階層的クラスタリングされたデータ集合 $X=\{x_1,x_2,\ldots,x_n\}$（$x_i \in R^d$）を考える．ラベル L を X の k 個の部分集合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．

$$C_{ij}=\left\{\begin{array}{ll}1 & x_i\&x_j(\text{belong same cluster})\\ 0 & \text{otherwis}\end{array}\right\}.$$

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

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

内積 $\langle L_1,L_2\rangle$ は，コーシー・シュワルツの定理 $\langle L_1\rangle,\langle L_2\rangle \leq \sqrt{\langle L_1,L_1\rangle\langle L_2,L_2\rangle}$ 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．

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

この相関測度を実際に用いた例として，Ben-Hur らの手法[7] があげられる．この手法は，元のデータ集合 W から，要素数が 50%より大きい部分集合 W_1，W_2（$|W_1|=|W_2|$）をランダムに作成し，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 $W_1\cap W_2$ に含まれる要素に注目する．樹形図を

$2 \sim |W_1| - 1$ 個のクラスタ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要素が W_1 と W_2 の間で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する．この操作を繰り返して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分析結果を得ることができる．

部分集合の共通部分に対して相関測度を用いる手法における安定性は，「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，異なる部分集合に対して要素を統計的に取得し，繰り返し同じ処理をほどこさ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

本章では，2.2 節であげた手法のような，統計的手法によら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）樹形図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）そのうえで要素を 1 個追加し，階層構造の変化を検出する，があげられる．

3.1 クラスタ間距離と安定性

ここでは，クラスタ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，一部の要素の変化による要素間あるいはクラスタ間の距離関係の逆転が考えられる．この距離関係の逆転の起こりやすさは，樹形図上に表されているクラスタ間距離にも依存するが，距離だけでは判定できない．つまり，樹形図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．

3.2 仮想要素追加法による安定性のモデル化

階層構造が変化する要因となりうる要素の変化には，以下の 3 通りが考えられる．

(1)　要素の増加

(2)　要素の減少

(3)　要素の値の変動

従来の安定性測度には，このうちの (2) または (3) が用いられている．しかし，一定量の要素の削除や，全要素の値の変動のために，ランダムサンプリングならびに統計的手法が必要である．

そこで本手法では，上記 (1) のケースに着目する．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見なす．いま，**図 1** (a)(1) のような 3 クラスタ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，たとえば図 1 (a)(2) 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，階層構造は図 1 (a)(3) に示すように図 1 (a)(1) と変化していない．これに対して，図 1 (a)(2) 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図 1 (a)(5) 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分かる．

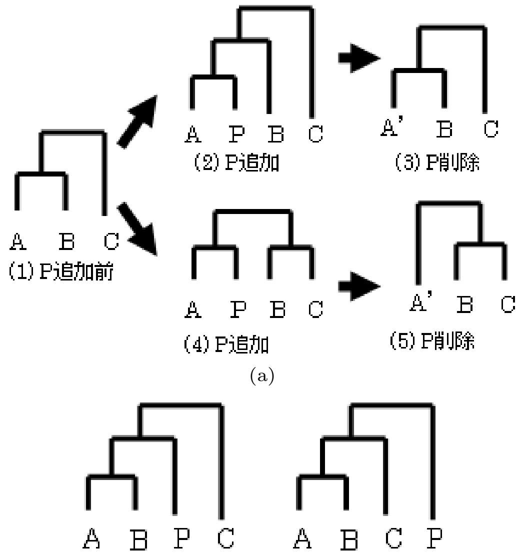


図 1　仮想要素 P の追加削除による階層構造変化
Fig. 1　Hirarchical structure changing by adding and deleting temporary element P.

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．

3.3 階層安定度の定義

本節では，前節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，データ要素が存在する n 次元空間内に要素 P を追加した場合，P が A，B，C いずれか 1 つのクラスタ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_a とする．たとえば，図 1 (a)(2)，(4) 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_a に含まれる．一方，図 1 (b) の 2 つの例は，いずれも P は A，B，C の少なくとも 2 つが結合したクラスタと結合している．このような場合，階層構造変化は起こりえないため，そのような P の値は対象領域 R_a からは除外する．

領域 R_a 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_a に占める R_s の領域の大きさの割合，すなわち R_s/R_a を，A，B，C の 3 クラスタからなるクラスタの階層安定度と定義する．領域の大きさは，原則として n 次元ユークリッド空間の超体積で測る．ただし，クラスタリングにおける距離尺度のとり方によっては，n 次元超体積では求められないこともありうる．そのような場合の対処の例を 3.5 節で述べる．

3.4　2 次元ユークリッド空間における適用例

前節で述べた安定度を，2 次元データに適用した例を示す．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によるものとする．

3.4.1　3 要素間での階層構造変化の例

ここでは，簡単のためにまず 3 クラスタがすべて 1 要素からなる場合に，追加要素が階層構造変化を引き起こす様子を解析する．

まず，3 要素 A，B，C の配置として，各要素間距離が以下の 2 種類の場合を考える．

(a)　$|\mathrm{AB}| : |\mathrm{AC}| = 1 : \sqrt{2}$

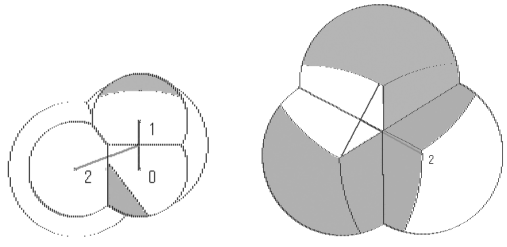
(b)　$|\mathrm{AB}| \approx |\mathrm{BC}| \approx |\mathrm{CA}|$

それぞれにおいて，3 要素の近傍に追加要素 1 個を置き，これを動かしたときの階層構造変化の様子を，**図 2** に示す．図中の直線や曲線は，要素 P を含めた階層構造が変化する境界線である．また，網掛け部分は，要素 P によって本質的な階層構造変化が引き起こされる部分を示す．

図 2 (a) では網掛け面積が小さく，図 2 (b) では網掛け面積が大きくなっている．これは，データ間距離の差が小さい場合，つまり図 2 (b) のような場合には階層構造の逆転が起こりやすく，逆に図 2 (a) のようにデータ間距離の差が大きい場合には逆転が起こりにくいことから理解できる．また，|AB| に対して |AC|，|BC| が十分大きい場合には階層構造の変化は起こらない．このように，本質的階層構造変化の起こる領域の面積が，クラスタの安定度を示す指標となりうることが分かる．

3.4.2　境界線の構成要素

前項で示した階層構造変化の境界線について，理論的に解析する．これらの境界線は次の要素から構成される．



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

図 2　構造変化が生じる領域：追加要素が網掛け位置に追加されたときに構造変化が生じる（括弧内は安定度）

Fig. 2　The regions cause structure changing. When a temporary element added in hatched area, structure changing is caused.

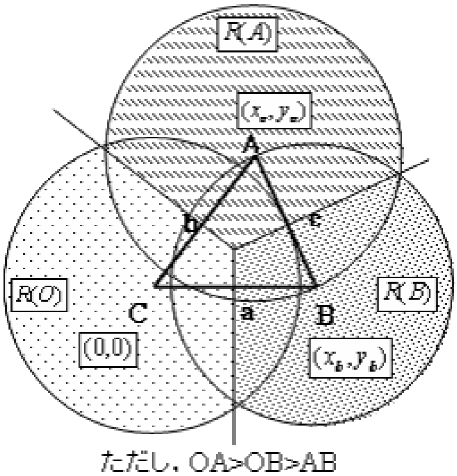


図 3　座標の定義と領域 R_a

Fig. 3　Definition of coodinates and region R_a.

太線：　3 要素のいずれか 1 つが，追加した要素と直接統合する境界

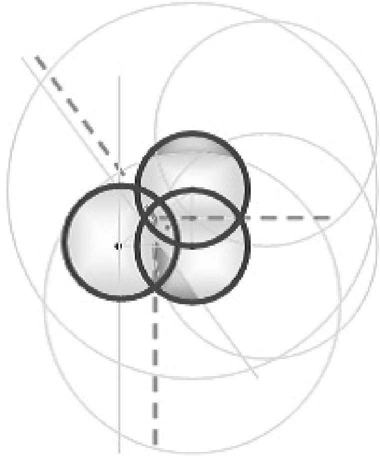
太破線：　3 要素間のボロノイ線

細線：　3 要素のいずれかと追加した要素とが統合したとき，他のデータとの距離の大小関係が変化する境界

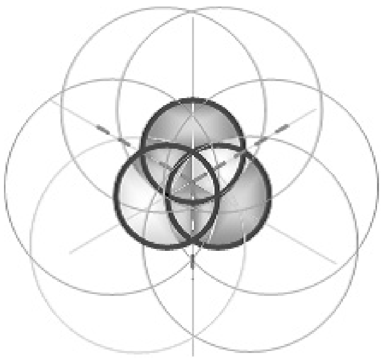
前項の配置（図 2 (a)，(b)）におけるこれらの境界線を，**図 3**，**図 4** に示す．

以下，これらの境界線の方程式を示す．ただし，座標は図 3 のとおりとする．図中および下記数式の a，b，c はそれぞれ |BC|，|CA|，|AB| を表す．まず，2 要素 A，B が互いに結合する以前に，追加要素が 3 要素 A，B，C のいずれかが結合するための条件は，追加要素が以下の円内部に存在することである．

A を中心とする円：



(a) $|AB| : |AC| = 1 : \sqrt{2}$



(b) $|AB| \approx |BC| \approx |CA|$

図 4　階層構造変化が起こる領域と起こらない領域の境界線：濃く網掛けされた部分で階層構造変化が起こる

Fig. 4　Borderlines between the region R_a. Deepen hatched areas are the regions cause.

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

B を中心とする円：

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

C を中心とする円：

$$x^2 + y^2 = c^2$$

これらの内部の領域が R_a となる．さらに，3 本のボロノイ線により領域 R_a は 3 個の領域に分割される．

AC 間のボロノイ線：

$$x_a x + y_a y = b^2/2$$

BC 間のボロノイ線：

$$x_a x + y_a y = b^2/2$$

AB 間のボロノイ線：

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

分割された領域をそれぞれ $R(A)$，$R(B)$，$R(C)$ とする．これらの領域は，追加要素がどの要素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．次に，追加要素が最初のステップで A，B，C と結合したとき，そのクラスタの重心をそれぞれ A′，B′，C′ とおく．領域 $R(A)$，$R(B)$，$R(C)$ の内部に追加要素が入るとき，階層構造が変化するための条件について，それぞれの領域ごとに以下に述べる．なお，各境界線は次のルールで命名する．

Z_{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，C の距離関係である．$R(A)$ 内部で A′，B，C の距離関係の境界線は，

円：$B_{A'C}$

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

円：$C_{A'B}$

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

直線：$C_{A'B}$

$$x_a x + y_a y = a^2 - x_a x_b - y_a y_b$$

であり，データ追加に関係して階層構造の変化が起こる条件は，

- 円 $B_{A'C}$ の外側
  または
- 直線 $C_{A'B}$ で二分される領域のうちの C 側

となる．

(ii)　$R(B)$

以下同様に，

円：$A_{B'C}$

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

円：$C_{B'A}$

$$(x + x_b)^2 + (y + y_b)^2 = 4a^2$$

直線：B'_{AC}

$$x_a x + y_a y = b^2 - x_a x_b - y_a y_b$$

この場合の変化が起こる条件は，

- 円 $A_{B'C}$ の外側
  または
- 直線 B'_{AC} で二分される領域のうちの C 側

となる．

(iii)　$R(C)$

同様に，

円：$A_{BC'}$

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

円：$B_{C'A}$

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

直線：C'_{AB}

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

変化が起こる条件として，

- $A_{BC'}$ の内側
  または

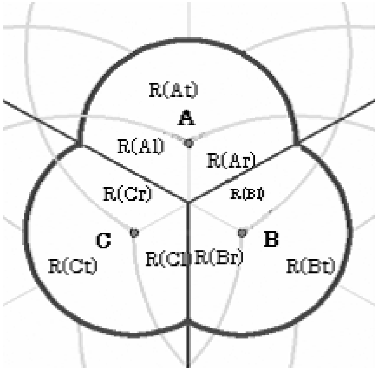


図 5　境界線で分割された R_a のラベル
Fig. 5　The labels of R_a divided by borderlines.

- $B_{C'A}$ の内側

となる．

3.4.3　要素間の階層安定度

要素間距離が |AB| = |AC| = |BC| である場合について，境界線によって分割される各領域に**図 5** のように名前をつける．これらの 9 領域が Ra となる．このうち，階層構造が変化する領域 Ru は $R(At)$，$R(Ar)$，$R(Bt)$，$R(Bl)$，$R(Cl)$，$R(Cr)$ である．

要素間距離が異なる場合は，これらの領域のすべてが存在するとは限らないので，存在する領域について面積を求めればよい．たとえば，図 2 (a) は Ru として $R(At)$ と $R(Cr)$ のみが存在する場合である．

階層安定度が最も低くなるのは，|AB| = |AC| = |BC| のときである．この場合，$R(At)$，$R(Bt)$，$R(Ct)$ 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ので，安定度は 1/3 となる．よって階層安定度の値域は $1/3 \leq$ 階層安定度 ≤ 1 となる．

仮想要素追加法による安定度を，以下のように近似的に計算する．R_a 領域を描画し，その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．図 2 の網掛けした領域が R_u である．このときの安定度は 0.88 であり，(b) では最も不安定なときの理論値 1/3 に近い 0.34 となる．

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．結果を**図 6** に示す．不安定な領域を目立たせるため，安定度が低くなるほど白くなるように色付けしてある．ここで，円内の黒領域は 3 個目の要素が A または B と先に結合するため，対象外であることを示す．3 要素間の距離がほぼ等しくなる

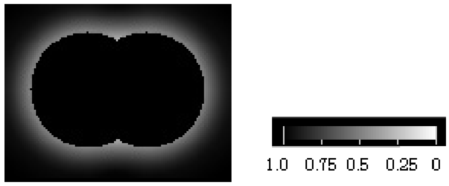
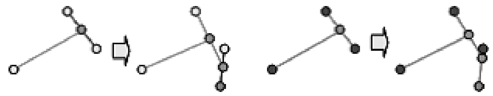


図 6　2 要素を固定し，3 要素目を 2 固定要素の周りで動かしたときの安定度分布．白いほど不安定である．ただし，中央の黒い円を 2 つ重ねた領域は安定度計算対象外の領域である
Fig. 6　Stability distribution when 2 elements are fixed and the third element is moved around fixed 2 elements. If whiter, more fragile. However, the region in 2 black circle is not object for stability calculatation.



(a) 要素の場合（1 : 1）　　(b) クラスタの場合（n : 1）

図 7　結合後のクラスタ代表値
Fig. 7　Centroid of a cluster after combined.

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．

3.4.4　クラスタ間の階層安定度

3.4.1 項では簡単のため要素数 1 のクラスタを対象としていたが，ここで一般化し，複数個の要素を持つクラスタでの階層安定度の求め方について述べる．

ここで例として用いている重心法では，クラスタを結合する際にその重みを考慮して新たな代表値を計算する．そこで本手法でもクラスタに適用する際には，結合先のクラスタの重みを考慮する必要がある．クラスタに対して仮想要素が結合することを考える．クラスタは所属する要素の数だけの重みを持っているため，仮想要素とクラスタの結合では，代表値は重みの分だけ，クラスタ側に寄ることになる（**図 7**）．これには重心法の距離計算をそのまま用いればよい．クラスタ A のデータ数を n としたとき，(x_1, x_2) の仮想要素と代表値 (a_1, a_2) のクラスタ A から構成される新たなクラスタ A′ の代表値 (a'_1, a'_2) は，次のようになる．

$$a'_1 = \frac{na_1 + x_1}{n+1}, \quad a'_2 = \frac{na_2 + x_2}{n+1},$$

これにともなって 3.4.2 項の境界線の式も変化する．

3.5　コサイン距離の適用

本節では，現実の多次元データのクラスタリングで多く用いられる，コサイン距離，群平均法への適用について述べる．

コサイン距離は，2 要素 $\mathbf{a_i}$，$\mathbf{a_j}$ の距離 l_{ij} を下記の式で求める．

$$l_{ij} = \frac{\mathbf{a_i} \cdot \mathbf{a_j}}{|\mathbf{a_i}||\mathbf{a_j}|} = \hat{\mathbf{a}}_\mathbf{i} \cdot \hat{\mathbf{a}}_\mathbf{j},$$

ただし，$\hat{\mathbf{a}}_\mathbf{i}$，$\hat{\mathbf{a}}_\mathbf{j}$ はベクトル $\mathbf{a_i}$，$\mathbf{a_j}$ を正規化したものである．

群平均法は，クラスタ間の距離を，両クラスタの全要素間で可能なすべての対で求めた要素間距離を平均することで求める．要素数がそれぞれ n_A，n_B であるクラスタ $A = \{\mathbf{a_i}\}$，$B = \{\mathbf{b_i}\}$ のクラスタ間距離は次の式で求められる．

$$d_{AB} = \frac{1}{n_A + n_B} \sum_i \sum_j \hat{\mathbf{a}}_\mathbf{i} \cdot \hat{\mathbf{a}}_\mathbf{j}. \qquad (1)$$

3.5.1 安定度の考え方

安定度は 3.3 節で述べたとおり，領域 R_a に占める R_u の割合として定義する．

コサイン距離では，式 (1) に示すように，要素ベクトルの方向だけに依存し，大きさ（原点からの距離）は無関係である．したがって，領域 R_a は n 次元空間上で原点を頂点とする錘状の無限領域となり，3.3 節で述べたように領域の大きさを n 次元超体積として求めることはできない．

そこで，すべての要素ベクトルを正規化して考える．このとき，n 次元データであれば，正規化した要素は n 次元の単位超球面上に分布する（**図 8**）．領域 R_a などの大きさは，この超球面上の $(n-1)$ 次元超体積として求める．

3.5.2 安定度の求め方

多次元データの場合，安定度を幾何学的に求めることは困難である．ここでは，次の手順で単位超球面上で R_a 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サンプリング点を配置し，各点が Ra 内の点であるかを調べる．

(2) R_a 内の点である場合，この点を要素に追加したときのクラスタ階層構造変化の有無を調べ，R_s か R_u かを判定する．

(3) サンプリング点の個数比から，安定度 R_s/R_a を求める．

3.5.3 計算の高速化

サンプリングで求める場合，データの次元が高くなるほど計算量は指数的に増大する．ここでは，計算の高速化について述べる．

クラスタ $A = \{a_i\}$，$B = b_i$，$C = c_i$ において，それぞれの要素数を，n_A，n_B，n_C とおき，それぞれのクラスタについて要素正規化したものの平均をそれぞれのクラスタの代表点として，次のように定義する．

$$\bar{\mathbf{a}} = \frac{1}{n_A} \sum \hat{\mathbf{a}}_\mathbf{i},\ \bar{\mathbf{b}} = \frac{1}{n_B} \sum \hat{\mathbf{b}}_\mathbf{i},\ \bar{\mathbf{c}} = \frac{1}{n_C} \sum \hat{\mathbf{c}}_\mathbf{i},$$

このとき，安定度は内積 $\bar{\mathbf{a}} \cdot \bar{\mathbf{b}}$，$\bar{\mathbf{b}} \cdot \bar{\mathbf{c}}$，$\bar{\mathbf{c}} \cdot \bar{\mathbf{a}}$，および，$|\bar{\mathbf{a}}|$，$|\bar{\mathbf{b}}|$，$|\bar{\mathbf{c}}|$，$n_A$，$n_B$，$n_C$ の関数となる．そこで，前処理として，これらのパラメータ個々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，表引きと補間により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. クラスタ要素の広がり度合いの可視化

3 章で述べた階層安定度は，2 個のクラスタを結合したときに，他の 1 個のクラスタとの間に不安定な状態になるかどうかの尺度を与える．ところが，各クラスタの要素の広がりは必ずしも反映されていないため，要素が大きく広がり，クラスタを分離すべきでない場合（たとえば**図 9**）にも高い安定性を示すことがある．そこで，本章では，クラスタ要素の広がり度合いを可視化する手法について提案する．

1 個のクラスタを 2 個に分離できるかどうかを，要素の広がり度合いから判別するために，次のような手順で可視化を行う．

(1) 分離後の 2 個のクラスタの各代表値を算出．

(2) 所属するすべての要素を，2 つの代表値を結ぶ直線上に投影．

(3) 投影した結果を樹形図上にクラスタごとに高さを変えて可視化．

注目クラスタの代表値を F，対向クラスタの代表値

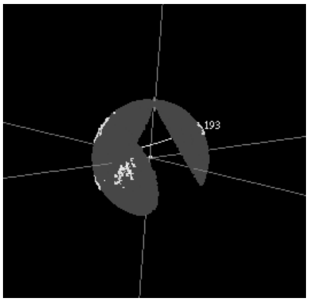


図 8 3 次元データに適用した R_a の例
Fig. 8 An example of R_a applied into 3D data.

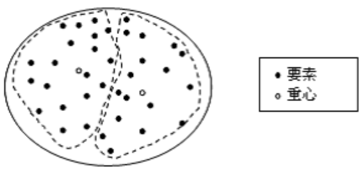
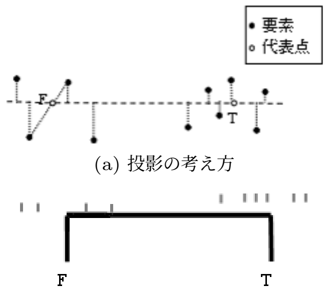


図 9 クラスタ要素が十分広がっているため，分割すべきでないクラスタの例
Fig. 9 An example of the cluster should not to be devided because cluster elements are enoughly spreaded.



(a) 投影の考え方

(b) 上記の要素の樹形図への可視化イメージ

図 10　要素の広がり度合いの表現
Fig. 10　Expression of density of the elements.

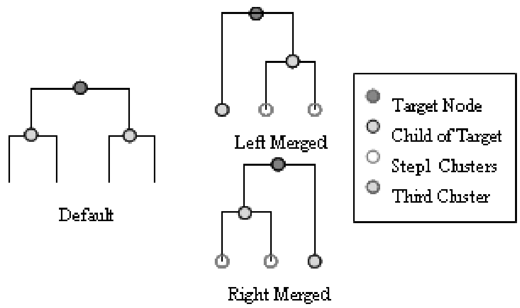


図 11　安定度計算対象ノードの選択
Fig. 11　Node selecting that is object for calculation of stability.

をT，注目クラスタの要素をEとすると，クラスタ要素の投影E' は次式で表される．

$$\overrightarrow{\mathrm{FE}'} = \frac{(\overrightarrow{\mathrm{FE}} \cdot \overrightarrow{\mathrm{FT}})}{|\overrightarrow{\mathrm{FT}}|^2}\overrightarrow{\mathrm{FT}}.$$

樹形図では，縦軸にクラスタの距離をとったとき，結合される2個のクラスタを縦棒で描画し，それらを横棒でつなぎ合わせることで結合を表現する．提案手法では，縦棒の上端を代表点とし，横棒上およびその延長に広がり度合いを描画する．

図 10 (a) に，クラスタが大きく2個に分かれるデータ集合の場合の例をあげる．

上記の操作をすべての要素で繰り返すことで，クラスタ全体の要素の広がり度合いが表現できる．

5. 階層安定度および広がり度合いの可視化

本章では，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．これにより，階層安定度と要素の広がり度合いを考慮したクラスタ分割数の決定が可能となる．

5.1　樹形図への適用

樹形図は二分木であるので，最下層から2階層以上，上にあるノードは3または4ノードからなる．

提案手法は3要素に対する安定度計算法であるので，樹形図に適用するにあたっていずれかの3ノードを選ばなくてはならない．すなわち**図 11** Default のうち左右どちらの子から孫ノードを用いるか定める必要がある．ここではクラスタリングの際の結合順序に従い，左側が先に結合している場合にはLeftMergedを，右側が先に結合している場合にはRightMergedを選択する．このように選んだ3ノードで安定度を計算することで，クラスタリング後の階層構造は各ノードが安定度という値を持つ二分木となる．

5.2　各階層の安定度の可視化

この安定度を樹形図に表現するために，**図 12** の表現を用いる．この表現では選ばれている3ノードを明示するため，対象ノードを結んでできた三角形を安定度に割り当てた色で塗りつぶす．

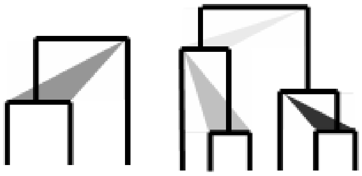


図 12　安定度の表現
Fig. 12　Expression of stability.

この手法を用いて25個の要素からなるデータ集合を階層的クラスタリングし，樹形図に安定度を付加した結果を，**図 13** に示す．安定度は明度にマッピングしている．図13では，背景が黒で描かれているため，不安定な階層を目立たせるため，不安定なほど背景の対向色である白に近づくようにしている．

5.3　要素の広がり度合いも含めた可視化

前節の安定度可視化手法に加え，4章で提案した広がり度合いの可視化手法を適用した例を**図 14** に示す．これにより，クラスタの構造，階層安定度，要素の広がり度合いが一度に把握できる．

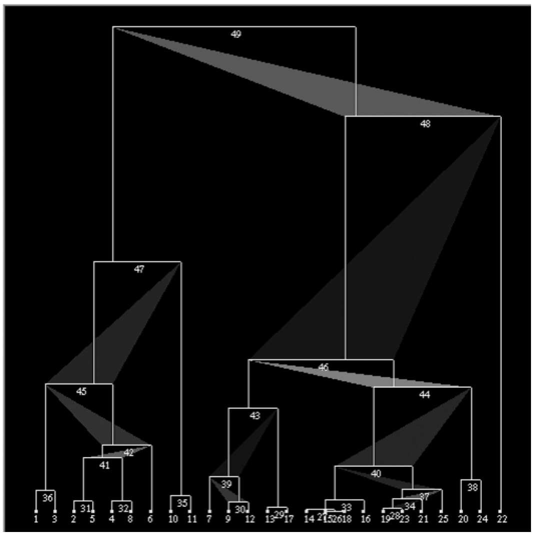
5.4　クラスタ分割数の決定

提案手法による可視化を用いた場合，クラスタの分割の可否は次のように判断できる．

(1)　あるクラスタにおいて，右と左の要素分布が分離していないとき，2クラスタへの分割は不可．
(2)　あるクラスタにおいて，それを構成する3個のサブクラスタ間での安定度が低いとき，2クラスタへの分割は不可．

さらに，クラスタ対の一方に比べて他方の要素数が極端に少ない場合は，少ないほうのクラスタは特異点（クラスタ）と考えられ，分割しないほうがよい．

例として，図14に示す3次元データを対象とし，



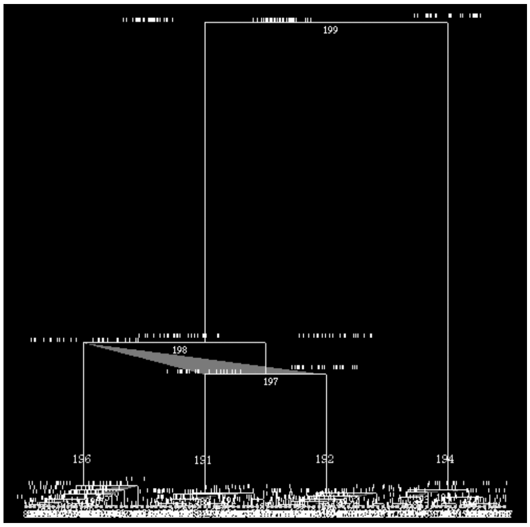
(a) 可視化結果

(b) 安定度スケール

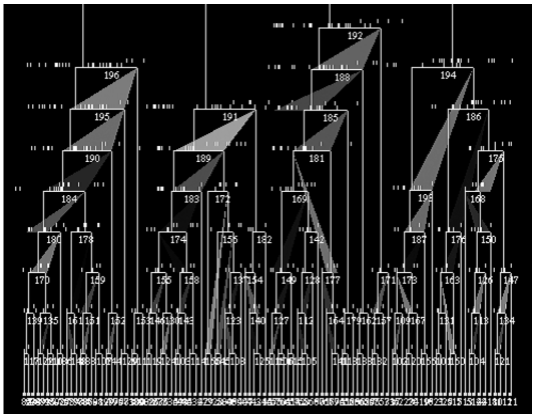
図 13 階層安定度の可視化結果（25data）
Fig. 13 Result of stability visualization.

要素間距離をコサイン距離，クラスタリング法を群平均法としたときのクラスタ分割数を求める．図 14 中に表記されている数字はクラスタ番号である．まず図 14 (a) の 199 を見ると，要素分布が右と左で完全に分離している．また，その下に三角形が示されていないことから，2 クラスタに分離しても安定である．したがって，198 と 194 への分割は望ましいと考えられる．次に，198 を見ると，その下に明度の高い三角形が見られる．このとき，196，191，192 の 3 つのクラスタのうち，191 と 192 を結合して 197 を作ることが不安定であることを示している．したがって，194，196，197 の 3 個のクラスタへの分割は望ましくない．197 は，三角形，要素分布の両面から，分割することが望ましい．197 より下層に焦点を当てた可視化結果を図 14 (b) に示す．これを見ると，196，191，188 の層はすべて要素分布が分離していないので，分割をしないほうが望ましいことが分かる．したがって，分割数は 194 と 198 の 2 分割か，193，196，189，195 の 4 分割が適当である．図 14 (c) に元となった 3 次元データを示す．

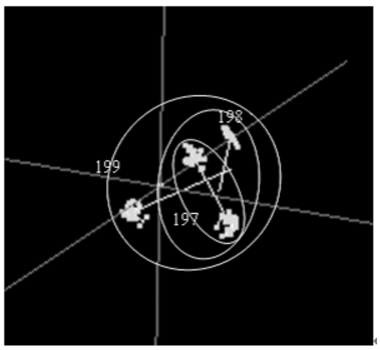
なお，最終的な分割数の決定は，個々のアプリケーションやデータ特性に依存するため，ユーザによって判断されるべきである．このことは，Ben-Hur らの実験[7] においても，クラスタ分割数を実験結果に加えてデータ特性を加味した結果から得ていることからもいえる．



(a) 樹形図の全体像



(b) 197 より下層の拡大表示



(c) サンプルデータ集合（100data）

図 14 要素の広がり度合いの可視化結果（100data）
Fig. 14 Result of density visualization.

5.5 広がりの可視化が適用できないケース

前節の手法では，クラスタが n 次元空間内でおおむね等方的に広がっていることを想定している．した

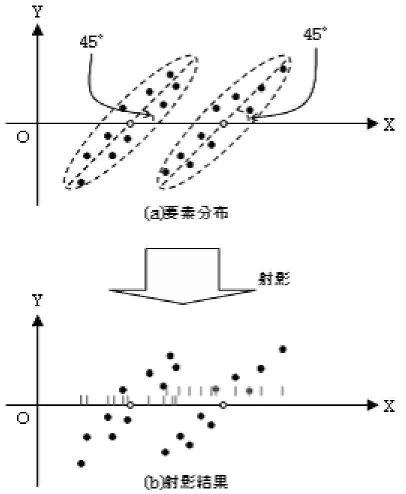


図 15 分離を正しく可視化できないケース．クラスタは分離されているが，投影すると一部が重なってしまい，分離しているように見えない

Fig. 15 A case that is not be able to visualize correctly. 2 clusters can't look to separate, because when project, parts of both of cluster are overlaped.

がって，そうでないケースでは，正しく分割できないことがありうる．たとえば，**図 15** のような場合を考える．図 15 (a) は，xy 平面上に 2 つのクラスタが分布している例である．距離尺度は，ユークリッド距離，重心法で考える．このとき，2 つのクラスタの代表値はともに x 軸上に位置している．また，それぞれの要素は，x 軸に対して 45 度に傾いた長軸を持つ楕円形状に分布している．このとき，代表点を結ぶ直線上（x 軸上）にそれぞれの要素を射影すると，図 15 (b) のような可視化結果が得られるが，この可視化結果からは 2 つのクラスタが分離している様子は分からない．このような場合，5.4 節の分割方法ではクラスタを正しく分割することができない．

6. 既存手法との比較実験と考察

本章では既存手法として 2 章でも述べた Ben-Hur らの手法[7] を取り上げて計算機上に実装し，提案手法の有効性，問題点を検証する．

6.1 実験概要

(1) 目的：提案手法の有効性，問題点を検証するために，Ben-Hur 法と提案手法の実行結果を取得．

(2) 内容：提案手法と Ben-Hur 法を計算機上に実装し，同一のデータを適用して実行結果を比較．

(3) 使用データ：100 点の 3 次元空間データ．あらかじめクラスタ数が 4 つになるように作成したもので，5 章で使用したものと同じデータ．

(4) 取得データ：両者とも次のデータを取得する．

(a) プログラムの実行時間

(b) プログラムを実行して得られたクラスタ数および結果の妥当性

(5) クラスタリングアルゴリズム：提案手法，Ben-Hur 法ともに要素間距離をコサイン距離，クラスタリング法に群平均法を使用．

(6) Ben-Hur 法の実行条件：

(a) データサンプリング率：80%

(b) データ比較回数：100 回

(c) クラスタ数判定方法：実行結果から得られたデータを安定度でソートしたものを表計算ソフトに入力して累積百分率の折れ線グラフを作成し，十分安定で数が最多のクラスタ数を取得．

(7) 提案手法の実行条件：

(a) サンプリング法により計算．

(b) クラスタ分割数は，Ben-Hur 法と同じく最多のものを採用．

(8) 実験環境

(a) CPU：Pentium4 3.2 GHz

(b) RAM：1 GB

(c) 実行環境：cygwin Xserver 6.8.99.901-4

表 1 実験結果

Table 1 Result of experiment.

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

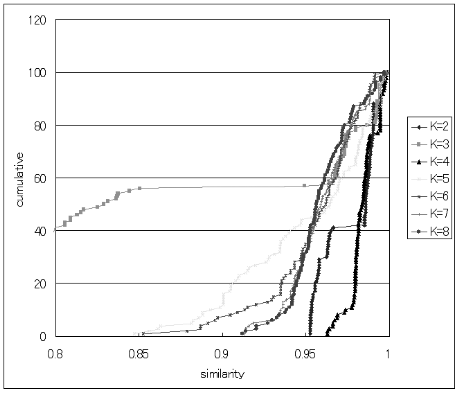


図 16 Ben-Hur 法の結果

Fig. 16 Result of Ben-Hur method.

6.2 実験結果

上記のとおり実験したところ，**表 1** のような結果が得られた．Ben-Hur 法の実行結果を**図 16** に示す．このグラフは，横軸に安定度，縦軸に累積度をとり，実行した 100 回の比較で得られた安定度を昇順にソート

し，クラスタ数ごとに累積させたものである．グラフの凡例にある K は，クラスタ数である．図 16 を見ると分かるとおり，グラフの立ち上がりが最も急激，すなわち全体的に安定度の高い $K=2$ および $K=4$ が候補となるが，$K=2$ は，$K=4$ に次いで安定であるが，最多のクラスタを分割数とするため，結果は $K=4$ とする．提案手法の実験経緯と可視化結果は，5.4 節でも見たとおり，望ましい分割数の候補は 2 または 4 である．最多クラスタ数を採用すれば，こちらも分割数は 4 となる．

6.3　プログラム実行時間

提案手法は，今回はサンプリング法を用いているためかなり時間がかかっている．しかし，3.5 節で述べたように，表引きと補間により，高速化できると考えられる．一方，Ben-Hur 法は，部分集合を統計的に調べる手法であるため，十分な試行回数が必須であり，高速化は困難である．また，要素数が増えると，処理時間も増大する．

6.4　結果の分かりやすさ

Ben-Hur 法においては，クラスタ分割数は，図 16 のようなグラフから読み取ることができる．しかし，樹形図との対応がとれないため，データ集合が実際にどのように分割できるかは不明である．したがって，Ben-Hur 法では，樹形図を求めたクラスタ数になる距離で切ることで個々のクラスタを得るしかない．それに対して提案手法では，樹形図上に分割の判断に必要な情報が提示されるので，データ集合がどのように分割されるかの様子が把握できる．また，分割に際しては，樹形図の階層は固定されることなく，特定のクラスタを深く分割することも可能である．

6.5　少数データへの適用

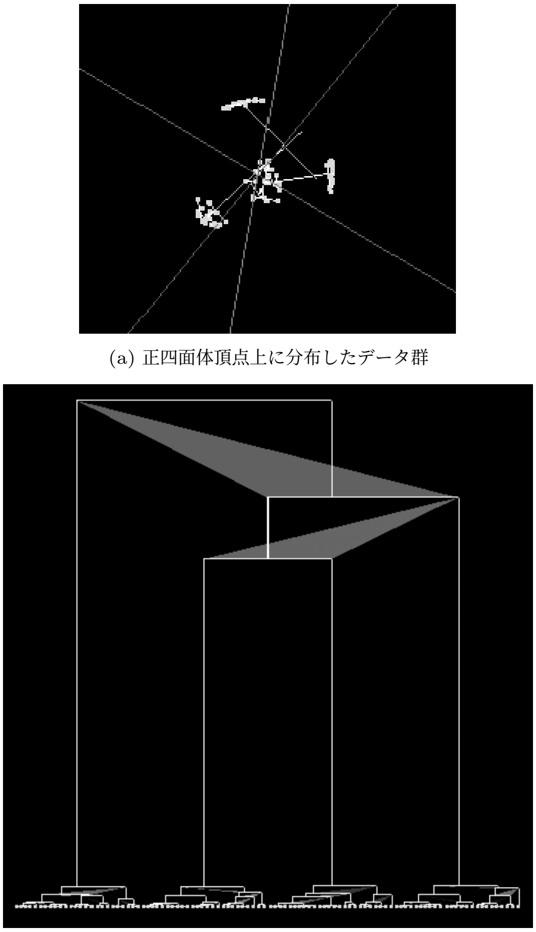
本手法の優位点としてデータ数が少ない場合に適用できることもあげられる．Ben-Hur 法では，データ数が一定以上ない場合には十分な信頼性を得られないのに対し，提案手法はわずか 3 個の要素からなるクラスタに対しても，安定度を計算できる．

6.6　本手法での安定度可視化の限界

ここでは提案手法の特性から，その限界について述べる．

提案手法では，図 1 上で，P の追加によって A，B，C がそれ以外のクラスタと先に結合する場合については考慮していない．そのような場合，上の階層の構造が大幅に変化し，解析が複雑化するためである．その分だけ，安定度の算出と可視化が不正確になる可能性がある．これについては今後の課題とする．

たとえば**図 17** (a) のような，データが正四面体の頂点付近に分布している例を考える．このとき，データ群それぞれの代表点はほぼ正四面体の頂点に位置するので，各々のクラスタ間距離はほぼ等しくなる．そのため，データの微小変動によってこれらのクラスタがどのような順番にでも結合しうる．実際にクラスタリングし，安定度を計算した結果を図 17 (b) に示す．提案手法では注目している 3 要素が不安定であることを示すことができるが，注目しているもの以外の要素を含めた結合順序が入れ替わる可能性を示すことはできない．たとえば，**図 18** の可能性 (1)，(2) は可視化結果から読み取ることができるが，(3) については読み取ることができない．



(a) 正四面体頂点上に分布したデータ群

(b) 安定度可視化結果

図 17　提案手法における例外ケース
Fig. 17　Example of excepted case in proposed method.

7. 結　　言

本研究では，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，階層安定度に加えて要素の

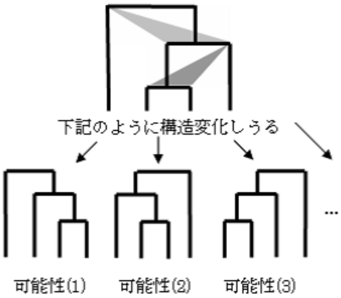


図 18　階層構造変化の可能性
Fig. 18　Possibility of hierarchical structure changing.

広がり度合いを可視化することで人間の直感に沿っ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，より人間の直感に沿ったクラスタ分割ができる．今後の課題として，まず計算時間の効率化・高速化があげられる．次に，提案手法は様々なクラスタリングアルゴリズムに対して適用できるため，今後は現実の大規模なアプリケーションに適用し，有効性を検証することもあげられる．また，6.6 節で述べたような，注目する 3 クラスタ以外と先に結合するようなケースへの対応も課題としてあげられる．

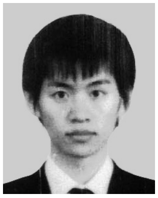
謝辞　本研究の一部は，科学研究費補助金（萌芽 17650024）の援助を受けている．

参 考 文 献

1) Jain, A.K., Murty, M.N. and Flynn, P.J.: Data clustering: A review, *ACM Computing Surveys*, Vol.31, No.3, pp.264–323 (1999).
2) Ranghavan, V.V. and Ip, M.Y.L.: Techniques for measuring the stability of clustering: A comparative study, *ACM SIGIR 1982*, pp.209–237 (1982).
3) Rand, W.M.: Objective criteria for the evaluation of clustering, *Journal of American Statistical Association*, Vol.66, No.336, pp.846–850 (1971).
4) Corneil, D.G. and Woodward, M.E.: 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) Yu, C.T.: 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) Fowlkes, E.B. and Mallows, C.L.: A method for comparing two hierarchical clusterings, *Journal of the American Statistical Association*, Vol.78, No.78, pp.553–584 (1983).
7) Ben-Hur, A., Elisseeff, A. and Guyon, I.: A stability based method for discovering structure in clustered data, *Pacific Symposium on Biocomputing*, Vol.7, pp.6–17 (2002).

(平成 19 年 2 月 2 日受付)
(平成 19 年 3 月 23 日再受付)
(平成 19 年 5 月 14 日採録)

渡部　秀文（学生会員）
平成 14 年東京農工大学大学院工学研究科電子情報工学専攻博士前期課程修了．同年（株）NTT データ入社．IDC 業務および請求関連システム開発に従事．平成 18 年より東京農工大学大学院生物システム応用科学府生物システム応用科学専攻博士後期課程在籍.



南雲　　拓
平成 18 年東京農工大学大学院生物システム応用科学府生物システム応用科学専攻博士前期課程修了．同年（株）リコー入社.

一宮　和正（学生会員）
平成 19 年東京農工大学工学部情報コミュニケーション工学科卒業．同年 4 月より東京農工大学大学院工学府情報工学専攻博士前期課程在籍.

斎藤　隆文（正会員）
昭和 62 年東京大学大学院情報工学専攻博士課程満期退学（平成 2 年修了）．同年日本電信電話（株）NTT 研究所勤務．平成 3〜4 年米国ブリガムヤング大学客員研究員．平成 9 年東京農工大学工学部助教授，平成 14 年より同大学院生物システム応用科学府教授．CG，映像処理，可視化等の研究に従事．工学博士.

宮村（中村）浩子（正会員）
平成 16 年お茶の水女子大学大学院人間文化研究科博士後期課程修了．同年 4 月東京農工大学大学院生物システム応用科学府助手，平成 19 年より同大学院助教．主にボリュームビジュアリゼーション，インフォメーションビジュアリゼーションの研究に従事．博士（理学）．