Vol.2009-CG-135 No.6
2009/7/13

仮想要素追加法を用いた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[†] 一宮 和正[††,☆] 斎藤 隆文[†] 古谷 雅理[††]

本報告では，新しい階層的クラスタリングの安定性モデルと，その可視化手法について提案する．階層的クラスタリングでは，データにわずかなノイズや誤差が混入した時にその結果が大きく影響を受けることがある．このような問題は安定性の問題として，安定なデータ分割を得るための多くの研究がなされている．
そこで，本報告では，従来法で見られるクラスタリング結果全体ではなく，それぞれの階層ごとに安定度を定義し，それを樹形図上に可視化することで，従来法より具体的で安定かつ高速なクラスタ分割数を得る手法を提案する．

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

Hidefumi Watanabe, [†] Kazumasa Ichimiya, [††,☆]
Takahumi Saito[†] and Tadasuke Furuya[††]

We propose a new mathematical model for analyzing the stability of hierarchical clustering results. In addition, we also present a method to visualize stability calculated by the new stability model. Hierarchical clustering has a problem that the result often changes according to a little difference of input data like the noise and the error, etc.
In our methods, we calculate the stability of hierarchical clustering by each hierarchy of cluster structure, but whole cluster structure like existing methods. Furthermore, we visualize stability on the dendrogram. We then show the method to decide number of division of clusters more concrete, reliable, and faster than existing methods.

1. 緒言

本報告では，仮想要素追加法を用いた階層的クラスタリングの安定性の計算モデルと，その可視化手法を提案する．

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ことを目的とする．階層的クラスタリングは，バイオインフォマティクスやマーケティング, 文書分類等の分野において利用されている．

階層的クラスタリングは，純粋に数学的な手法であるため，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．

階層的クラスタリングの安定性に関する研究は多く存在する．しかし，その多くは安定で最適な分類数を得ることにとどまっており，クラスタ構造を十分に分析できる情報が得られないことがある．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)が渡部らによって提案された．本報告では，仮想要素追加法の概要と，距離尺度ごとの幾何構造について述べる．また，実装の高速化手法として，特定次元による任意次元の近似手法と，表引きによる手法について述べる．さらに，階層的クラスタリングの出力結果である距離付樹形図に，仮想要素追加法で計算した階層安定度を樹形図に可視化する手法について述べる．

2. 先行研究

本節では，一般的な階層的クラスタリングの可視化手法について述べる．また，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．

2.1 一般的な階層的クラスタリングの可視化

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（図 1）が使われるが，そこから得られる情報は少ない．しかも，扱われるデータは 100 を超えるような多次元になることも多い．そのため，単純な手法ではグラフィカルな形でデータ属性や相関を提示できない．そこで，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3)．また，データ

[†]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 Richo Company, Ltd.

Vol.2009-CG-135 No.6
2009/7/13

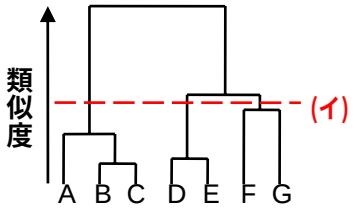


図 1　距離付樹形図

が大規模になり，樹形図が 1 画面に収まらない時の圧縮表示手法について提案している．さらに，データを 2 次元座標にマッピングした結果や，異なる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している．

2.2 Ben-Hur らの手法

Ben-Hur らは，E.B.Fowlkes らの類似性測度4)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，ヒストグラムと分割数ごとの安定度が読み取れる散布図により可視化している．

Ben-Hur らの手法では，安定性の測定結果から分割を得るときに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，ヒストグラムと散布図から分割数 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，結果の信頼性を保つために統計的な繰り返し処理が必要となる．また，ヒストグラムや散布図により安定性を可視化しているが，これらは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，Ben-Hur らの手法では，得られた分割数と実際に得られるデータ分割が厳密には対応せず，さらなる分析が必要となる．

3. 仮想要素追加法

本節では，クラスタの階層構造そのものに着目して各階層の安定度を求めるモデルである，仮想要素追加法 1)について述べる．仮想要素追加法の適用法として，階層的クラスタリングでよく利用される距離尺度であるユークリッド距離と，コサイン距離での手法を紹介する．また，距離尺度ごとの多次元での近似手法，さらに計算の高速化について提案する．

3.1 仮想要素追加法の概要

仮想要素追加法では，図 2(a)のような，2 つのクラスタ要素にもう 1 つのクラスタ要素が結合した単純な階層構造を計算単位とする．対象となる要素には，階層構造の最下層のリーフ要素をはじめ，リーフ要素やクラスタが結合した上層のクラスタが含まれる．上層のクラスタが計算対象となる場合，クラスタの代表点が実際の計算対象となる．これに仮想的なクラスタ要素を 1 つ追加し，クラスタリングを行ったときの構造変化に着目する．例えば，図 2 (a)のような階層構造に仮想要素 P を追加してクラスタリングを行った場合，図 2 (b)や(c)などの結果が得られる．図 2 (b)，(c)ともに追加した P が A とクラスタを形成しているが，特に図 2 (c)では，B は A より先に C とクラスタを形成している．このように，P 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の起こりやすさから安定性を測定する手法について述べる．まず，計算対象 3 要素のうち，最初に結合する 2 つを A，B，残りを C とする．追加する仮想要素 P は，A，B，C のいずれかと先に結合する場合だけを対象として考える．例えば，図 2(d)のような時，A，B，C いずれか 2 つが結合した後で P が結合している．このようなときは，階層構造変化は起こり得ないため，対象から除外する．P が対象となるのは，P が A, B, C のいずれかから|AB|以内の距離にある場合に限られる．これを満たすデータ領域を R_a とする．図 3 に領域 R_a の例を 2 つ示す．P を領域 R_a 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_a 中で，実際に階層構造に変化が起こる領域を R_u，階層構造変化が起こらない領域を R_s とする．図 3 では R_s は白抜きの領域，R_u は着色された領域となる．このようにして求められた領域 R_a で，R_a 全体に占める領域 R_s の比率を計算し，この 3 点による階層構造の階層安定度とする．この安定度は最も不安定な場合で 1/3，安定な場合で 1 となる．図 3(a)では階層安定度は約 0.88，図 3(b)では 0.34 となる．

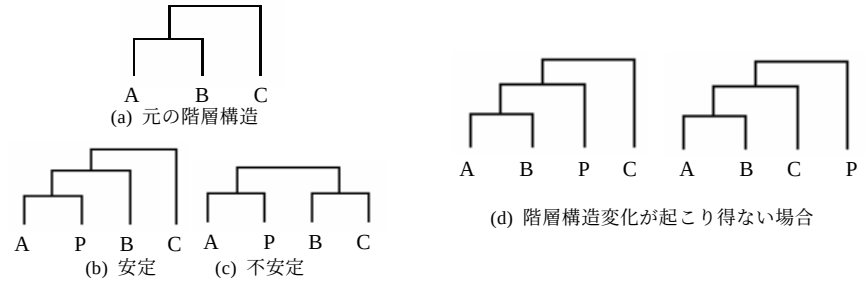
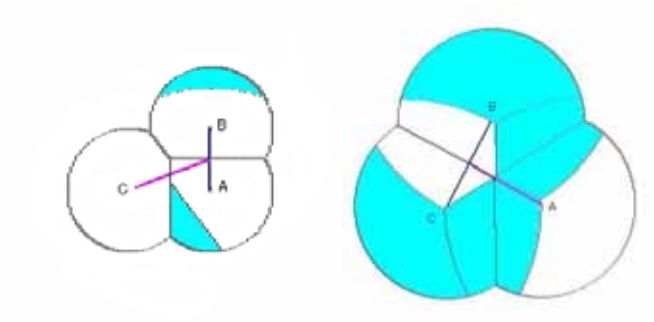


図 2　階層構造の変化

Vol.2009-CG-135 No.6
2009/7/13



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

図 3　領域 R_a の例．括弧内は階層安定度を示す

仮想要素追加法の特徴として，上記のように階層安定度を数値で求められることがあげられる．また，Ben-Hur の手法のような，試行の繰り返しによる統計的な処理が必要ないことも特徴としてあげられる．

3.2 特徴空間の幾何構造と安定度の算出法

本項では，階層的クラスタリングで用いられる距離尺度のうち，ユークリッド距離による手法と，コサイン距離による手法に関して，安定度算出の考え方を幾何構造の面から述べる．対象間の距離は，絶対距離であるユークリッド距離でとることが人間の直感に合うことが多い．しかし，解析対象のデータの特性によっては相対距離をとる方が対象の特性をよく反映した分割を得られることがある．たとえば，DNA マイクロアレイ解析では，非線型なコサイン距離をとることが一般的である．

(1) ユークリッド距離の場合

ユークリッド距離を用いる場合，領域 R_a を考える際の特徴空間は線型となる．図 3 はユークリッド距離を用いた場合の R_a である．特徴空間が 2 次元であるため，R_a は 2 次元平面上に円が 3 つ結合した形で現れる．3 次元であれば，R_a は A，B，C を中心とした，半径|AB|の球が 3 つ結合した領域（図 4）となり，n 次元であれば n 次元超球が 3 つ結合した領域となる．

仮想要素追加法の計算機への実装に関して，文献 1)ではサンプリング法による近似計算手法が提案されている．まず，R_a 内を適切な密度のサンプル点により均等に走査する．次に，これらのサンプル点が R_s 内に置かれた個数を数え上げる．最後に，全サンプル点の個数に対する R_s 内となった点の個数の割合を求めれば，安定度を求めることができる．

しかし，サンプリング法の場合，R_a の大きさは n 次元空間では n 次元超体積となり，計算量は指数オーダになる．そのため，特徴空間の次元が 100 を超える階層的クラスタリングへの適用は現実的ではない．そこで，比較的高速に計算可能な 3 次元の特徴空間を走査することにより，4 次元以上の任意次元の安定度を近似的に求める手法を提案する（以下，高速化サンプリング法）．

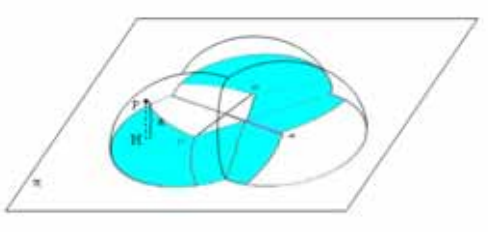
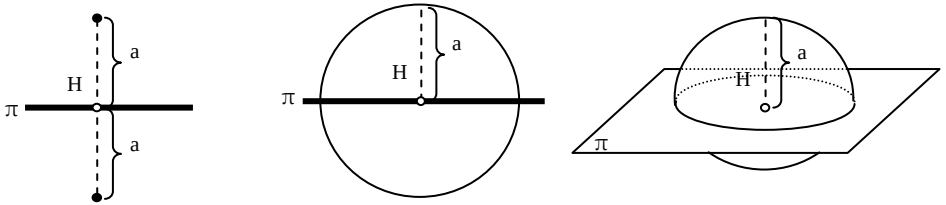


図 4　3 次元空間における R_a の概念図



(a) 3 次元：2 点　　(b) 4 次元：(4-2)次元超球面 →円弧　　(c) 5 次元：(5-2)次元超球面 →球面

図 5　H から距離 a を持つ点群のイメージ

R_a は A，B，C を中心とする 3 つの球によって構成される（図 4）．3 つの球の中心を通る平面をπとすると，R_a はπに対して上下対称となる．そこで，πの片側の領域だけを対象としてサンプリングを行えばよい．また，このときの仮想要素 P が R_s か否かは，P と A，B，C の距離関係により一意に決定される．3 次元において，R_a はπに対して上下対称であるため，このような P は R_a 内に 2 点存在する（図 5(a)）．

次に，4 次元以上で考える．平面πは 3 次元のときと同様に定義できる．P が R_s か否かは，3 次元のときと同様，P と A，B，C の距離関係により一意に決定される．P と同じ距離関係を持つ点は，4 次元以上では次元が高くなるため連続に存在する．このような点は，P からπにおろした垂線の長さを a とするとき，この垂線の足 H から距離 a である点の集合になる．このことから，n 次元の空間においては P と同じ距離関係を持つ点の集合は，H を中心とする半径 a の $n-2$ 次元超球面となる（図 5(b)(c)）．

上記のことから，n 次元（4 以上）の特徴空間での安定度は，3 次元の特徴空間を

Vol.2009-CG-135 No.6
2009/7/13

サンプリングし，P とπの距離が a であるとき， R_u と R_s の個数を，重み $W^n(a)$ をかけて積算すれば近似できる．重み $W^n(a)$ は，半径 a の $n-2$ 次元超球面の表面積であり，ガンマ関数Γを用いて次のように表すことができる．

$$W^n(a) = \frac{2\pi^{\frac{n-2}{2}}}{\Gamma(\frac{n-2}{2})} a^{n-3}. \qquad (1)$$

(2) コサイン距離の場合

コサイン距離では，対象間の距離を，適当な共通の始点から対象を終点としたベクトルによって定義する．対象ベクトルを **A**，B とし，**A** と B のなす角を θ とすると，コサイン距離 $d(\mathbf{A},\mathbf{B})$ は次のように表される．

$$d(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{|\mathbf{A}||\mathbf{B}|} = \cos\theta. \qquad (2)$$

上式のとおり，コサイン距離ではベクトルの方向だけが距離尺度に考慮され，ベクトルの大きさは考慮されない．このことから，ベクトルを正規化し単位ベクトルとして扱ってもよい．この時，ベクトルの終点は，始点を中心とした n 次元単位超球面上に分布する．領域 R_a はこの超球面上にクラスタの代表点ベクトル **A**，B，**C** を中心とする球面半径 θ の超球帽が3つ結合した形で表現される．R_a などの大きさは $n-1$ 次元超体積で定義できる．図 6 に 3 次元空間上の領域 R_a の例を示す．図 6 では，領域 R_a 全体を色分けせずに示してある．

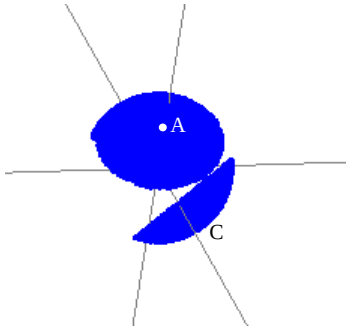


図 6　3 次元空間でコサイン距離をとったときの領域 R_a の例

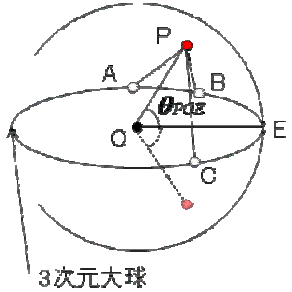


図 7　4 次元超球における各要素の配置

コサイン距離でもユークリッド距離と同様，サンプリング法では計算量が指数オーダになる問題がある．そこで，コサイン距離でも高速化サンプリング法を利用する．

A，B，C の位置関係は，それぞれの距離から 3 自由度で決定される．そのため，自由度が 4 である 4 次元超球面上では A，B，C は 3 次元大球（3 次元球面での大円に相当）上に配置することができる．コサイン距離での任意次元の近似は，まず A，B，C を 3 次元大球上に配置して考える．次に，仮想要素 P は 4 次元超球面上の領域 R_a 内でサンプル走査する．P が R_s か否かは，ユークリッド距離の時と同様に A，B，C と P の位置関係により一意に決定できる．4 次元空間では，この距離関係を持つ P は 3 次元大球をはさんで 2 つ存在する（図 7）．

次にこれを 5 次元以上で考える．3 次元大球は 4 次元の時と同様に定義できる．まず P が R_s か否かも，同様に A，B，C と P の位置関係により決定できる．次に，P と超球面の中心 O を結んだ直線を OP，OP を 3 次元大球上に射影した直線を OE とすると，P と同じ距離関係を持つ点群は，球面半径 θ_{POE}（図 7 参照）を持つ超球帽の弧となる．超球面の正規化半径が r の時，超球帽の弧の長さ $W^n(r)$ は， n 次元の空間においてはガンマ関数Γを用いて，次のように表せる．

$$W^n(r) = \frac{2\pi^{\frac{n-3}{2}}}{\Gamma(\frac{n-2}{2})} (r\sin\theta_{POE})^{n-4}. \qquad (3)$$

以降，ユークリッド距離の時と同様に R_u と R_s の個数を， $W^n(r)$ を重みとしてかけて積算すれば安定度が近似できる．

3.3 表引きによる高速化

仮想要素追加法の高速化手法として，表引き法について述べる．文献 1)では，2 次元ユークリッド距離，重心法での安定度分布に関して，計算対象の 3 個のクラスタ代

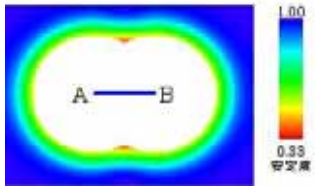


図 8　2 次元ユークリッド距離，重心法での安定度分布

表点のうち，A，B を固定し，C を動かすことで検証している（図 8）．

ユークリッド距離の場合，固定点間の距離が増加しても安定度分布は相似形を保つ．このことから，安定度は，3 クラスタの相対距離で求められる．ルックアップテーブルは，高速化サンプリング法を用いれば任意次元のものを高速に作成することができる．安定度を計算する際には作成したテーブルを読みだして補間とともに計算すれば，任意次元の安定度はより高速に求めることができる．

4.　安定性の可視化

仮想要素追加法を用いて計算された階層安定度は，各々樹形図の対応する階層に可視化することで，階層構造全体の安定性を可視化できる．一宮5)らは，安定性の可視化法や，可視化を用いたクラスタ分割手法について提案している．ここでは，樹形図への安定度の可視化について概要を簡単に述べる．

樹形図上に階層安定度可視化するには，まずクラスタの結合順序から計算対象を選択する．樹形図が図 9 のような階層を持つ場合，A と B が先にクラスタを形成するため，A+B のクラスタと C，D で安定度を計算し可視化する．可視化には，対象を明示するための，対象を結んだ三角形を描画し，安定度に割り当てた色で塗りつぶす．安定度が低い場合は，C, D のいずれかが先に A+B と結合する可能性があることを示す．

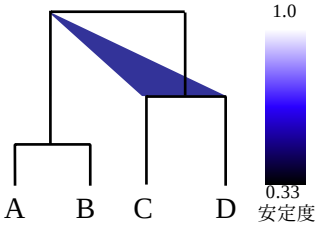


図 9　樹形図への階層安定度可視化例

5.　評価

本節では，3 節で述べた仮想要素追加法について，サンプリング法で計算される安定度の精度と，高速化サンプリング法および表引きを用いた高速化法の評価を行う．評価には，ユークリッド距離，重心法によるクラスタリングアルゴリズムを用いる．また，Ben-Hur の手法と計算速度の比較を行う．

5.1　サンプリング法の精度評価

3 節で述べたとおり，仮想要素追加法の実装には，サンプリング法の利用が基本となる．このサンプリング法の精度は，サンプリング密度に依存される．そこで，2 次元，3 次元について特徴空間の各次元のサンプリング間隔を変化させながら，計算される安定度の様子を検証した．安定度の値は，[1/3, 1]の範囲であり，4 節で述べたように，安定度は樹形図上に擬似カラーで可視化される．そのため，安定度の精度はユーザが相対的に大まかな目安がつけられる程度でよいと考える．ここでは，計算精度として，平均誤差で 0.01 以内，最大誤差で 0.05 以内を目指すこととする．

特徴空間に，安定度計算対象の 3 点のうち 2 点を固定し，3 点目をランダムに 1000 通り配置し，安定度を計算した．各次元のサンプリング点数は，5，10，20，40，80 の 5 段階とした．得られた安定度のうち，隣り合う 2 つの結果（5 と 10 など）の差分をとり，全体での平均，最大値，最小値を求めた（表 1）．これより，40～80 のときの差分平均が 2 次元で約 0.0015，最大値も約 0.0145，3 次元で 0.0008，0.0143 程度である．このことから，サンプリング点数は各次元 40 程度で十分に収束していると考えられ，理論値からの誤差もこの程度と考えられる．

5.2　サンプリング法と高速化サンプリング法の比較

次に，高速化サンプリング法について，精度と速度を評価する．高速化サンプリングのサンプリング空間の次元は，特徴空間の次元によらず 3 で固定される．そのため，対象のクラスタ 1 組あたりの計算量は，各次元あたりのサンプリング数の 3 乗に比例する．表 2 上 3 段は，クラスタ 1 組あたりの安定度計算時間と速度比をまとめたものである．これより，次元が高いほど大幅に高速化されていることがわかる．

表 1　2 次元及び 3 次元におけるサンプリング間隔ごとの安定度計算結果の差分
サンプリング点数の「a～b」の表記は，a と b で得た結果の差分をとることを示す．

	各次元のサンプリング点数	5～10	10～20	20～40	40～80
2 次元	差分平均	0.0513	0.01484	0.0046	0.0015
	差分最大値	0.5000	0.07425	0.0420	0.0145
	差分最小値	0.0000	0.00000	0.0000	0.0000
3 次元	差分平均	0.0496	0.01200	0.0030	0.0008
	差分最大値	0.4387	0.09230	0.0405	0.0143
	差分最小値	0.0000	0.00000	0.0000	0.0000

表 2　高速化サンプリング法の速度比較と精度検証結果

使用データ	4 次元	5 次元
実行時間・従来法[s]	199.9	12,810
実行時間・高速化法[s]	3.313	3.328
高速比[倍]	60.33	3849
誤差平均	0.00018	0.00024
誤差最大値	0.00655	0.00315
誤差最小値	0.00000	0.00000

計算精度については, 5.1 項と同様に 4 次元と 5 次元のランダムデータで評価した．サンプリング法と高速化サンプリング法でそれぞれ安定度を計算し，安定度の差分の平均値，最大値，最小値を計算した．4 次元データでは 1,000 点，5 次元データでは計算時間の都合上，250 点について評価した．結果を表 2 下 3 段に示す．5 次元での評価データ数は少ないものの，4 次元と同様に高精度に計算できていると考えられる．

5.3 表引き法による高速化の効果

表引き法の計算時間を，2 次元および 3 次元のデータを用いて評価し，結果を表 3 に示す．使用したデータのうち，データ(1)(2)は擬似的に作成したものであり，データ(3)は，UCI Machine Learning Repository 6)で公開されている"haberman data set"である．表 5 より，従来のサンプリング法と比較して，いずれのデータでも実行時間が大幅に高速化されていることがわかる．ただし，本評価では作成済みのテーブルを利用しているため，表 3 の計算時間にはテーブルの作成時間は含まれていない．

計算精度については，ここでも 3.2 項で使用したデータを表引きに適用し，従来手法と提案手法の安定度の差分から精度を検証した（表 4）．これにより，表引き法は十分使用に耐える精度であると考えられる．同様に 4 次元，5 次元でも高い精度で計算できている．

表 3　表引き法による高速化の評価

データ	(1)	(2)	(3)
要素数	22	37	306
次元数	2	3	3
実行時間・サンプリング[ms]	813	68,218	573,250
実行時間・表引き[ms]	78	94	141
表引き法のサンプリングからの速度比[倍]	10.43	725.7	4,065
実行時間(Ben-Hur の手法)[ms]	1,281	2,078	2,250
表引き法の，Ben-Hur の手法との速度比[倍]	16.42	22.10	15.95

表 4　高速化サンプリング法で作成した表引きによる安定度計算誤差の比較

使用データ	2 次元	3 次元	4 次元	5 次元
誤差平均	0.0007	0.0004	0.0004	0.0005
誤差最大値	0.0060	0.0080	0.0297	0.0378
誤差最小値	0.0000	0.0000	0.0000	0.0000

5.4 Ben-Hur の手法との速度比較

先行研究との比較として，Ben-Hur の手法との比較を行う．本項では，前項までと同じ環境で Ben-Hur の手法を実装し，表 3 のデータ(1)〜(3)に適用した．提案手法は高速化サンプリング法で作成された表引きを用いる. 両者の実行時間の合計を算出し，速度比を計測したところ，表 3 の下 2 段のようになった．これより，いずれの次元においても，提案手法の方が高速であることがわかる．

6. 結言

本報告では，仮想要素追加法を用いた階層的クラスタリングの安定性の計算モデルと，その可視化手法を提案した．仮想要素追加法を用いることで，従来法では考慮しなかったクラスタの階層構造についても安定度を定義でき，その可視化結果からより直観的で具体的なデータ分割を得ることができる．

今後の課題として，コサイン距離，群平均法を用いたクラスタリングでの高速化サンプリング法と表引き法の評価を，実際の利用シーンで利用されたデータをもとに行うことがあげられる．

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) J. Seo and B. Shneiderman, "Interactively Exploring Hierarchical Clustering Results," IEEE Computer Society Press, Vol.35, No.7, pp.80-86, 2002.
4) 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.
5)一宮和正, 渡部秀文, 宮村(中村)浩子, 古谷雅理, 斎藤隆文, “階層的クラスタリングの安定性の可視化,” 情報処理学会グラフィクスと CAD 研究報告 No.132 予稿集, pp.61-66, 2008.8.
6) UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/