グラフィクスとCAD　110－8
（２００３．２．１４）

木構造を持つ多変数データの可視化

鉢呂　重喜†　　斎藤　隆文††

本報告では，クラスタリングにより木構造化された多変数を持つデータに対し，樹形図による効果的な可視化手法を提案する．多次元の変数を持つ未知の大規模データを解析する際に，階層的クラスタリングがしばしば用いられる．対象データをクラスタ構造に従って一次元的にならべ，クラスタ構造を樹形図として可視化することにより，類似データ群の解析をサポートする．しかし，データの分布状況によって，類似したデータが異なるクラスタに分類されることが起こりうる．2002 年に，大野らによって類似データを極力近傍に配置する手線形配置手法が提案されている．本研究では，樹形図と共に種々の付加情報の提示を試みる．それによって，より直観的なデータマイニング支援を目指す．

Visualization of Multidimensional Data Set with Tree Structure

Shigeki Hachiro† and Takafumi Saito††

We propose visualization approaches for multidimensional data sets based on hierarchical clustering and dendrogram. To analyze multidimensional data sets, the hierarchical clustering method is commonly used. For visualizing the clustering result, dendrogram is usually used, where each member is linearly ordered depending on the cluster trees. However, the validness of the clustering depends on the data distribution, and similar members are sometimes classified in different clusters and placed far away. We cannot find such similarity from a dendrogram. Ohno et al. proposed an effective linear ordering where similar members tend to become adjacent. For more intuitive visualization, we propose to visualize additional information onto Ohno's dendrogram.

1. はじめに

1.1 背　　景

大規模データの解析を行う際に，データを木構造のような階層構造とみなし，絞り込みや分類を行うことがある．特に未知のデータに対して，階層的クラスタリングを行い，樹形図を生成し提示する可視化方法は，データを解析する一つの方法としてしばしば用いられる．しかし，一度階層的クラスタリングを行い樹形図として表示を行うと，樹形図の構造にとらわれてしまい，各要素間の関係といった重要な情報を見落としてしまうことがある．また，樹形図生成を行う場合，多変数のデータであっても線形配置を行う必要がある．

2001 年に大野らによって提案された，樹形図における線形配置法において，最近接端点接続法[1)] では，隣接する要素間の類似度の向上がなされている．また，ここでは付加情報の可視化によって，異クラスタに分散された類似要素の情報を提示している．

1.2 目　　的

本報告では，多変数データを，階層的クラスタリングをもちいて二分木構造化を行った際の可視化手法の提案を行う．線形配置手法は最近接端点接続法を用いる．特に，付加情報の提示により直観性の高い画像生成を目的とする．

1.3 実験データ

実験では，マイクロアレイ☆によって得られた遺伝子データと，アメダス☆☆によって得られた気温データを実験データとして用いる．

遺伝子のデータは，大腸癌症例における各遺伝子の control reference に対する発現比である．データの内訳は，原発巣肝転位なし 102 検体，原発巣肝転位あり 20 検体を含む大腸癌患者 122 検体と正常粘膜 12 検体，合わせて 134 検体に対して，それぞれ 801 種類の遺伝子データとなっている．各症例（大腸癌サンプル，正常粘膜）については，その発現比が 1 : 1 となるように median 補正を掛けている[2)]．

† 東京農工大学大学院　工学研究科　電子情報工学専攻
Division of Electronic and Information Engineering, Graduate School of Technology, Tokyo University of Agriculture and Technology
roppati@vc.base.tuat.ac.jp
†† 東京農工大学大学院　生物システム応用科学研究科
Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology
txsaito@cc.tuat.ac.jp

☆ スライドガラス上に数万個の DNA スポットを作成し，全ての遺伝子の動的挙動を効率的かつ定量的に計測する手法．

☆☆ 正式名称は，地域気象観測システム．全国約 1300ヵ所で局地的な気象現象を自動観測している．Automated Meteorological Data Acquisition System この頭文字をとって、AMeDAS と名づけられた．

アメダスのデータは，各観測点における 2001 年度における一日の最高気温と最低気温を 365 日分を用いる．ただし今回は，完全に対象データがそろっている観測点のみを入力対象とする．

2. クラスタリング

2.1 クラスタ間距離の定義

実験データを階層的クラスタリングにより，階層構造化する．実験では，遺伝子の類似度を用いた分類が必要であるため，クラスタ間距離は要素間の相関関係抽出係数を用いる必要がある．今回，相関関係抽出係数には Pearson's correlation coefficient を用いた[3]．

N 個の要素を持つ二つの遺伝子データ X，Y の相関関係係数 $R(X,Y)$ は次式で表すことができる．

$$R(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right)$$

ここで，

$$\Phi_G = \sqrt{\sum_{i=1}^{N}\frac{(G_i - G_{offset})^2}{N}}$$

このとき，G_i は要素 i における遺伝子 G の値であり，G_{offset} は観測値の分散，Φ_G は G の標準偏差である．また，対象データは control reference に対する発現比となっているため，データの値が 1.0 のとき，G_{offset} は 0 となる．ここで R の値は -1 から 1 の間に収まる，負の値が大きい程逆相関となり，対象データによって $R(X,Y)$ の絶対値が大きいほど，二つのクラスタの類似度は高いとする手法や，逆相関は相関が無いものとして扱う手法がある．今回，遺伝子データにおいては，$R(X,Y)$ の絶対値を相関として扱っている．アメダスデータでは，負の相関は見られなかったため，この手法による結果への影響は無い．

2.2 最近接端点接続法

クラスタリングの結果を樹形図として可視化する場合，各要素を線形配置する必要がある．階層的 クラスタリングでは，n 個の要素に対して必ず $n-1$ 個のクラスタが生成され，各クラスタはサブクラスタを二つ持つことになるため，2^{n-1} 種類の配置法が存在する．配置法次第では，類似しているはずの要素がかけ離れた位置に配置されていしまう可能性がある．樹形図として可視化する際には，極力類似要素同士が近くに配置されることが望ましい．

そこで，本研究では最近接端点接続法を用いる．最近端点接続法とは，次のような手法である．線形配列がすでに決定している二つのクラスタ $C2$ と $C3$ の両端の要素をそれぞれ，$e0$,$e2$，$e3$,$e5$ とする．二つのクラスタを結合する際に考えられる接合部分の隣接要素は，$e0-e3$, $e0-e5$, $e2-e3$, $e2-e5$, の 4 通り存在する．これらの中から，最も要素間距離が短いものを選択する．概要図を図 1 に示す．多変数データを扱っているため，完全な線形配置は不可能であるが，最近接端点接続法を用いると，隣接要素の類似度が向上する．

3. 樹形図の描画

最近接端点接続法を用いた階層的クラスタリングの結果を樹形図に出力した結果を図 2 に示す．既存の樹形図表示に比べ隣接要素の類似度が高いという点で勝っているが，データ解析を行うには不十分である．次に具体的な問題点を挙げ，効率的な樹形図の描画方法を提案する．

3.1 要素境界の表示

実験データのような大量のデータを一画面に表すと，樹形図が混みいってしまい，クラスタ領域を把握することが困難となる．図 2 の例では右端の二つの要素は，他の要素との類似度が低いため，最後までクラスタと結合せずに残っているが，この樹形図からその情報を読み取るのは困難である．一般的に，樹形図のラインを色分けするなどの方法が採用されているが，表示領域の問題やライン幅といった問題から，大量のデータを扱う場合に適切とは言えない．

そこで，樹形図描画部分の背景領域を活用し，クラスタごとの領域を明確にする．結果を図 3 に示す．樹形図に対して，半楕円を表示することで，クラスタの境界を直観的に表現している．図 2 における，右端の二つの要素のように特殊な場合でも図 3 では，容易に把握することができる．さらに，一つのクラスタに対して，四半楕円をつなぎ合わせた図形を適応させる結果を図 4 に示す．この図形の高さを類似度に対応させることで，線分による樹形図の情報を全て網羅することができ，さらに，図形部分に色づけを行うことで，クラスタ構造の深さを表現することができる．この表示方により，クラスタの境界を直観的に提示することができる．

3.2 拡 大 表 示

限られた表示領域の中に，多くの情報を盛り込むことは困難である．そこで，ズーム機能を設けることにより，データの細部の表示や，要素のオリジナルデータの表示が可能となる．図 5 に出力結果を示す．一つのクラスタに注目し，そのクラスタを中心として拡大表示を行う．拡大することで，より細かなクラスタ構造を表示し，要素のインデックスを表示するための領域を確保することができる．

3.3 選択的表示

ズームを行った場合，表示領域の確保は容易に行われるが，大域的な情報を完全に失ってしまうという欠点がある．具体的には，異なるクラスタに離散してしまった，類似要素の情報を失ってしまうことになる．

そこで，フィルタリング機能を設ける必要がある．ある一つの要素に注目し，その要素との類似度が低いデー

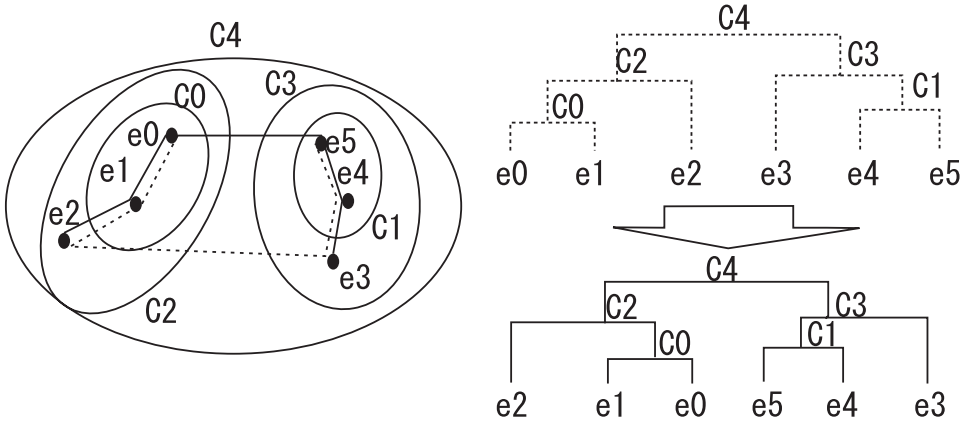

図 1　最近接端点接続法モデル図

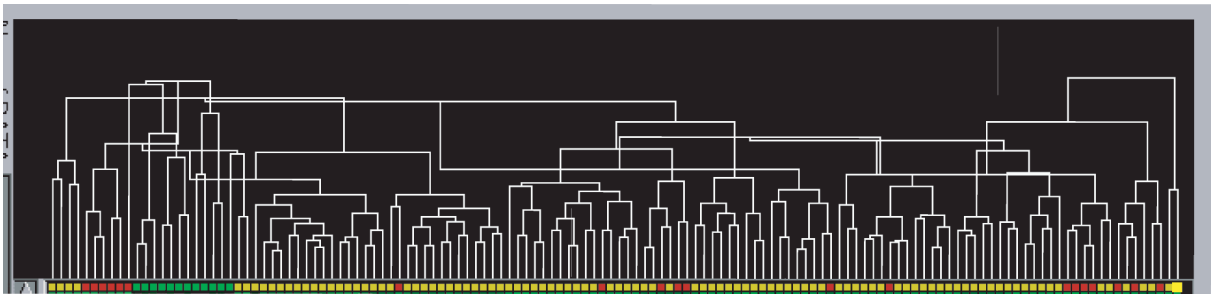

図 2　樹形図の描画

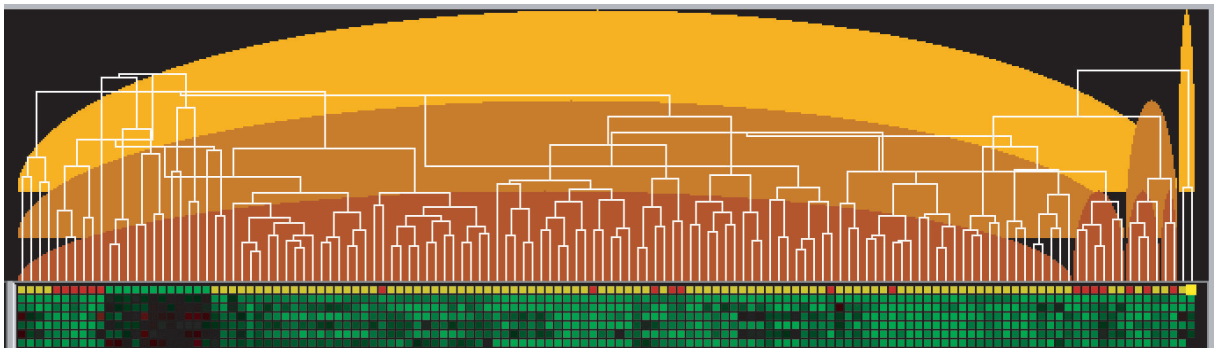

図 3　半楕円による境界表示

タを順番に切り落とす．出力結果を図 6 に示す．これにより，注目した要素と類似度の高い要素が残り，大域的なクラスタ構造を残したまま，詳細情報の表示領域も確保できるといった利点がある．

3.4 付加情報の可視化

一画面に納められるデータを増やし直観的なデータ解析を支援する方法を提案する．最近接端点接続法で線形配置されたデータに対して，付加情報を可視化する方法を考える．

(1) **類似度グラフ**

最近接端点接続法を用いても，類似要素が異クラスタに離散してしまうことは防げないため，類似要素が離れた位置に散在してしまう可能性がある．そこで，類似度情報を別途表示する必要がある．出力結果を図 7 に示す．

一つの要素をマウスで選択し，その要素に対する類似度を棒グラフで表示する．選択した要素に対するその他の要素の類似度や非類似似要素の特徴といった情報を直観的に得ることができる．単純な手法であるが，有用性が高い．

(2) **隣接要素間類似度グラフ**

隣接している要素でも類似度が低い場合が存在する．類似度グラフでは，一つの要素に着目した結果を出力しているが，選択したデータ以外の情報を読み取ることが困難である．

そこで，隣接する互いの要素間類似度を高さとした折れ線グラフを描画する．要素間類似度が高い場合は山，低い場合は谷となる．この手法により，異クラスタへ分かれてしまったデータでも隣接しているデータでも類似度の低いデータを見つけ出すことができる．結果画像を図 8 に示す．高低差の高い位置に着目することで，一度クラスタリングしたデータに対して，新たな分類を行うことができるが，これだけに注目して再分類を行うと，信頼性の面で危険がともなうため，なんらかの尺度を用いて，信頼性を向上させる必要がある．

4. 要素のインデックス情報表示

拡大表示や，選択的表示によって，要素のインデックスを樹形図内における表示が可能となっている．しかし，このままでは大域的な情報を表示した際に，各要素インデックスが不明である．マウスによる選択で，インデックス情報を表示させることも可能であるが，大域的な関係を重要視する場合に最適とは言えない．そこで，インデックスのテキスト情報を総覧できる表示が必要がとなる．最近接端点接続法により線形配置された要素のインデックスを表示する．ここで，表示されたインデックスと樹形図との対応が問題となる．当然スケールが異なるため，テキスト情報になんらかの情報を提示することで，樹形図との対応付けを行う必要がある．アメダスデータを用いたときの結果画像を図 9 に示す．

4.1 類似度表示

類似度グラフにおける要素間の類似度情報をフォントの輝度に割り当てることで，総覧を可能としている．樹形図部分において選択された注目要素の類似度を輝度の最大値として，各要素の注目要素との類似度が表現されている．

4.2 クラスタ情報の表示

直観的に樹形図部とインデックス表示部との対応付けを行うために，インデックス表示部にもクラスタ情報を提示する必要がある．特に，クラスタ境界部やクラスタ間の類似度といった情報が必要となる．

樹形図表示における四半楕円表示法の四半楕円の高さの総計を，グラフとして表示する．概念図を図 10 に示す．この表示法により，クラスタの境目やクラスタ間の類似度といった情報を，限られた表示領域の中で表現することができる．本手法は Cushion Treemap[4] において提案された表示法の応用といえる．

図 9 は，樹形図描画部分において SHINANOMACHI に注目した場合の例である．領域 a において，クラスタ情報の表示に着目すると HOTAKA と SAKUMA に大きなクラスタ境界があることがわかる．ここでフォントの輝度に着目すると，同様の部分で類似度が大きく変化しているのが読み取れる．この例では，クラスタリングが比較的うまくいっていることがわかる．

4.3 樹形図との対応情報

樹形図表示との対応をより明確にするために，インデックス表示部にスケールを樹形図に合わせた二つの矩形領域を表示する．図 9 の左図における右上方部に描画されている．領域 b の矩形領域には，クラスタ情報の表示で用いた四半楕円の高さの総計を輝度情報に直して，表示している．領域 c の矩形領域には，要素間の類似度を文字表示部分と同様の輝度で描画されている．また，中央部に樹形図表示部で選択した注目要素の位置が示されている．この表示法で，前述のクラスタ境界部に着目すると，クラスタの境界と類似度の変化がより直観的に読み取ることができる．

5. おわりに

実験データに対して，最近接端点接続法を利用した階層的クラスタリングを行い，木構造化した上で，樹形図表示とそれをより直観的に表示する手法の提案を行った．実装したシステムでは，インタラクティブな操作により，提案した様々な可視化手法を試すことが可能である．また，データの追跡を容易にするためにスムーズトランジションを導入した．

謝辞 株式会社日立製作所中央研究所メディカルシステム研究部バイオシステムセンタの神原秀記氏，大阪大学大学院病態制御外科の竹政伊知郎氏により遺伝子デーをご提供いただき，遺伝子データについての御教示いただきました．ここに深く感謝致します．

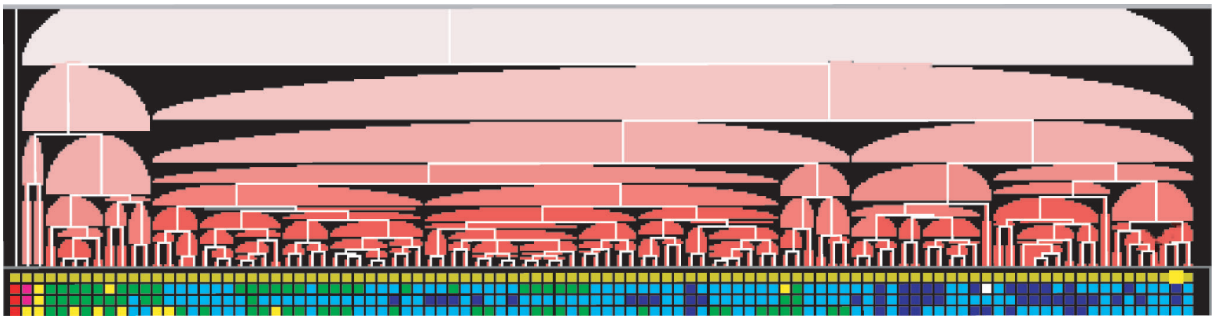

図 4　四半楕円による境界表示

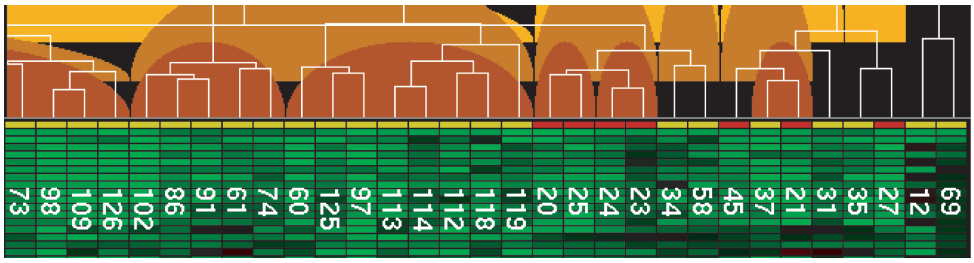

図 5　拡大表示表示

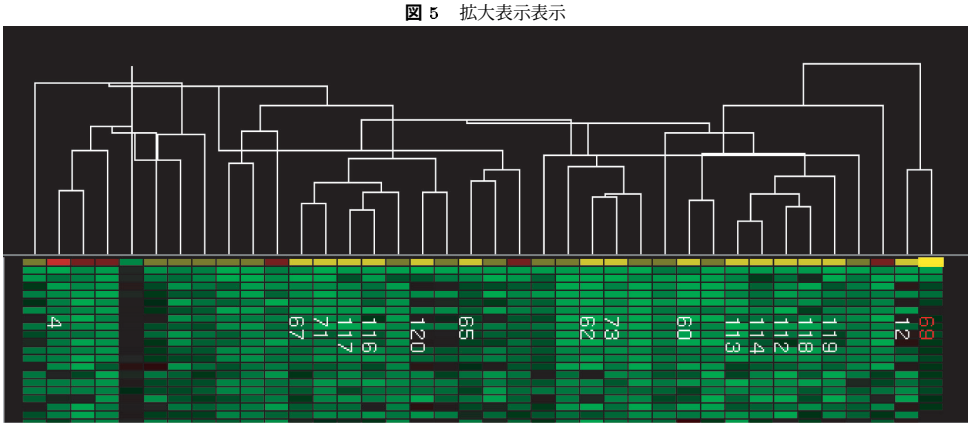

図 6　選択的表示

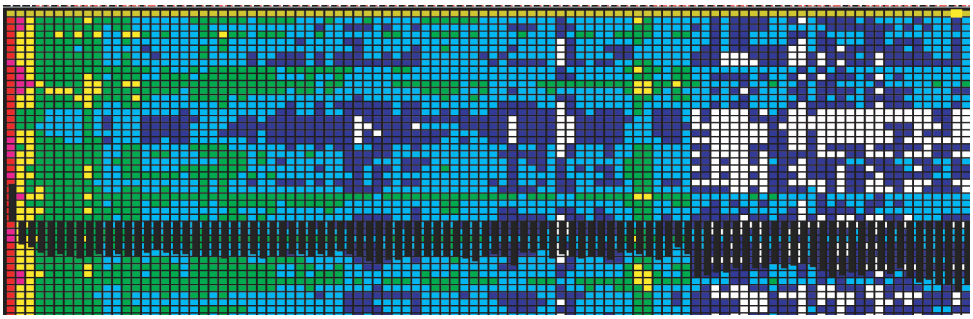

図 7　類似度グラフ

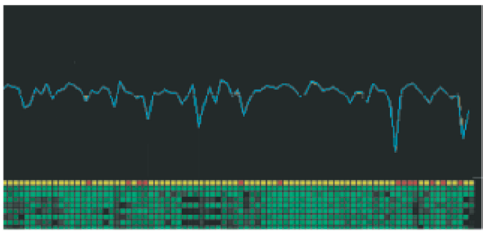

図 8　隣接要素間類似度グラフ

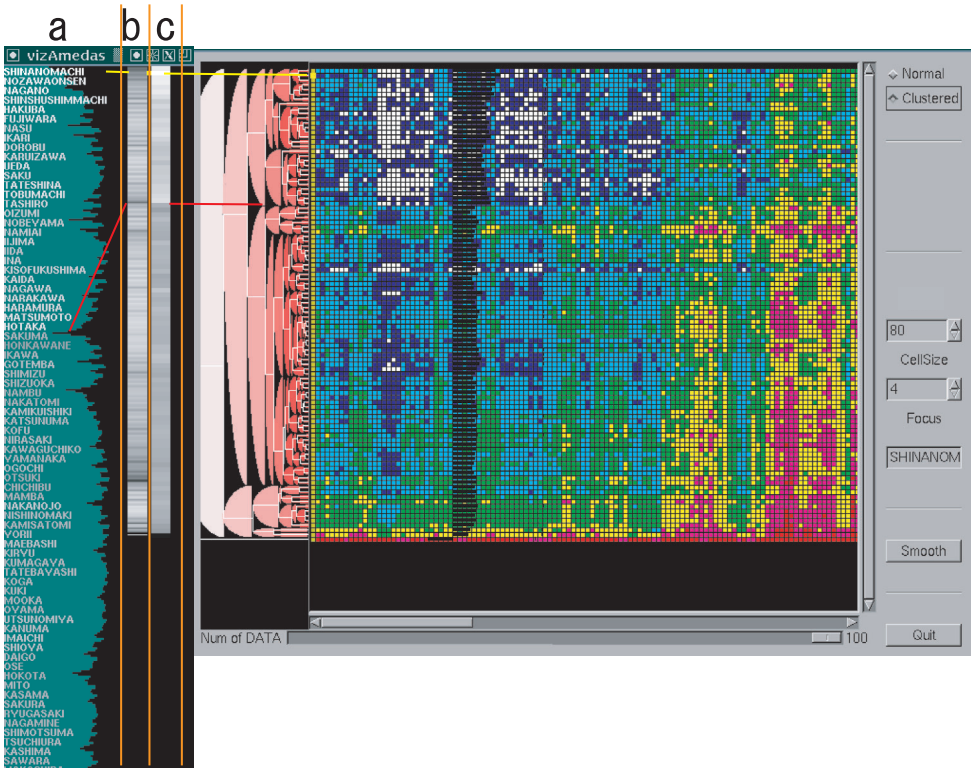

図 9　樹形図表示とインデックス情報表示

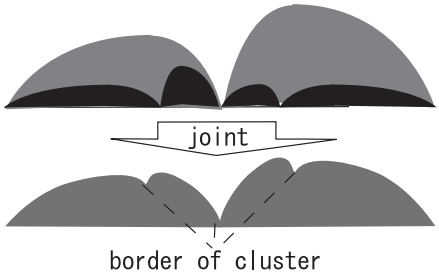

図 10　クラスタ構造表示概念図

参　考　文　献

1) 大野晋哉, 鉢呂重喜, 斎藤隆文, 階層的クラスタリングに基づく多次元データの可視化. 第 43 回情報処理学会プログラミングシンポジウム,pp.81-88,Jan.2002

2) I.Takemasa, H.Higuchi, H.Ymamoto,M.Sekimoto, N.Tomita,S.Nakamory,R.Matoba,M.Monden,and K. Matsubara, construction of preferential c-DNA microarray specialized for human colorectal carcinoma: Molecular shetch of colorectal cance. Biochemical and Biophysical Research Communication, Vol.285,pp.1244-1249,Jul.2001

3) Michael B. Eisen,Paul T. Spellman, Patrick O.Brown, and David Boststein, et al.,Cluster analysis and display of genome-wide expression patterns. Proc.Natl.Acad.Sci.USA,Vol.95,pp.14863-14868,Dec1998.

4) Jarke J. van Wijk, Huub van de Wetering IEEE Symposium on Information Visualization (INFOVIS '99), Cushion Treemaps: Visualization of Hierarchical Information pp.25-26,Oct.1999.