G-014

分割最適化クラスタリングの階層的可視化

Hierarchical Visualization of Partitional Optimization Clustering

桑原 俊† 森 康久仁† 松葉 育雄†
Shun Kuwabara Yasukuni Mori Ikuo Matsuba

1 はじめに

データマイニングは，大量のデータを解析することで有用な情報を抽出する技術である．その中でクラスタリングとは，与えられたデータから外的基準なしに部分集合に分割するようなデータ解析手法である．

近年，遺伝子という大量な情報に対して情報科学的なアプローチによって解析するバイオインフォマティクスが大変注目されている．バイオインフォマティクスでは，マイクロアレイという多数の遺伝子の発現状態を同時に観測する装置によって得られたデータをクラスタリングし，類似発現パターンを持つ群に分類することがよく行われる．それによって，遺伝子がどのような機能を持っているのかを推定することが重要な課題の一つとなっている．そこで本研究では，マイクロアレイデータから適切にクラスタリングを行い，情報を抽出するための可視化の方法論を提案する．

2 階層的クラスタリング

本研究では，与えられたデータに対してクラスタリングを行い，各クラスタを階層的に可視化することを目的としている．ここで，バイオインフォマティクスでよく用いられている階層的クラスタリングは，与えられたデータに対して，全データ間の類似度を求め，最も類似度が高い２つのデータをクラスタリングする．そして，構築したクラスタとその他のデータとの類似度を計算し，再び最も類似度が高い２つのデータをクラスタリングする．これを最終的に１つのクラスタとなるまで繰り返すという手法である．この手法の結果例を図１に示す．このように従来の階層的クラスタリングは，データを下層から１つ１つクラスタ化していき，最終的に１つのクラスタとなる凝縮型のクラスタリングといえる．そして，遺伝子発現の状態に応じた可視化方法 [1] によって視覚化する．

それに対して，本研究はデータに対して分割最適化クラスタリングを行い，それを階層的に可視化する．分

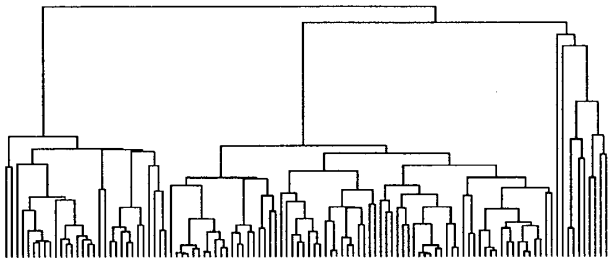

図 1: 従来手法の階層的クラスタリング結果の例

割最適化クラスタリングとは本来非階層的クラスタリングとも呼ばれ，任意の分割するクラスタの個数 k と分割の良さを示す評価関数を定め，その評価関数を最適にする分割を探索する手法である．しかし本研究では，データを任意の k 個に分割するだけでなく $2,3,\cdots,k$ 個に分割し，それら全ての場合においてクラスタ内のデータを可視化することによって分割最適化クラスタリングを階層的に図示する．よって，本研究で用いるクラスタリングは分割型のクラスタリングといえる．

3 使用データ

本研究では，マイクロアレイによる遺伝子発現データを用いた．マイクロアレイとは，ガラス等の小さな基板上に遺伝子の断片を高密度に配置 (アレイ) し，一度に数千から数万といった規模の遺伝子発現を同時に観察することができる装置のことである．このマイクロアレイによって得られたデータにおいて様々な解析を行うことにより，遺伝子間の相関関係や遺伝子の機能など様々な情報を推定することができる．

データの詳細としては，マイクロアレイによって得られた胞子形成に関わる遺伝子発現データで総遺伝子数 6118 個，時間点７個のデータ [2] を用いた．このデータは通常細胞の遺伝子発現量と，何か実験を施したサンプル細胞の遺伝子発現量を比較し，実験を施したことによって遺伝子の発現量が増えているのかもしくは減っているのかというのを対数比率で表し，それを時系列に並べたものである．このデータにおいて二つ以上の遺伝子を比較した時に，発現量の変化が類似して

†千葉大学大学院融合科学研究科

いるならばこれらの遺伝子の機能も類似しているのではないかと推測することができる．このように遺伝子発現データを解析し，クラスタリングを行うことで，いくつかのグループに分割することができ，それによって遺伝子の機能を予測することができる．

4 スペクトラルクラスタリング

本研究ではスペクトラルクラスタリング [3] を用いて遺伝子の分類を行っている．このクラスタリング手法は各データ間の類似度を評価関数に入力し，評価関数の最適解がある固有値問題の解に対応することを利用して分類を行う手法である．

ここで，W をデータ間の類似度行列，$e = (1,1,\cdots,1)^t$ としたとき $D = diag(We)$，q を N 次元ベクトルでデータがクラスタ A, B のどちらに属するのかを決定する要素とするとき，評価関数 J_m を (1) 式のように定める．

$$J_m = \frac{q^t(D-W)q}{q^tWq} = \frac{(W^{\frac{1}{2}}q)^t(W^{-\frac{1}{2}}DW^{-\frac{1}{2}}-I)(W^{\frac{1}{2}}q)}{(W^{\frac{1}{2}}q)^t(W^{\frac{1}{2}}q)} \quad (1)$$

このとき，

$$z = W^{\frac{1}{2}}q \quad , \quad X = W^{-\frac{1}{2}}DW^{-\frac{1}{2}} - I$$

とすると，(1) 式の最小値は行列 X の最小の固有値 λ_1 であり，それに対応する固有ベクトル z_1 によって実現される．しかし，$z_1 = W^{\frac{1}{2}}e, \lambda_1 = 0$ というのはクラスタ分割を成しえないため，2 番目に小さい固有値 λ_2 に対応する固有ベクトル z_2 を近似解として用いる．そして，この固有ベクトルの値から各データがクラスタ A，クラスタ B のどちらに属するのかを決定する．更に必要であれば，各クラスタ内の要素間の類似度の平均を求め，それが最も低いクラスタを同様の手法で更に分割していく．

また，本研究では類似度としてユークリッド距離の二乗平均の逆数を用いてスペクトラルクラスタリングを行った．

5 結果と考察

スペクトラルクラスタリングによって分割したクラスタ内の遺伝子発現データを可視化したものを図 2 に示す．この図は発現量の値が正であるほどより緑に，負であるほどより赤にプロットしたものである．各クラスタを見てみると，分割されたクラスタに含まれる遺伝子は非常に発現の仕方が類似していることが分かる．また，他のクラスタと比較してみるとクラスタごとに異なる発現の特徴を持つ遺伝子群に分割されているのが見てとれる．これらのことより，遺伝子発現の特徴を捉え，クラスタリングを行えたと考えられる．また，分割した各クラスタ内の遺伝子発現を階層的に図示することによってどのような特徴をもった遺伝子がどのようにクラスタリングされていくのかを視覚的に捉えることができるようになった．

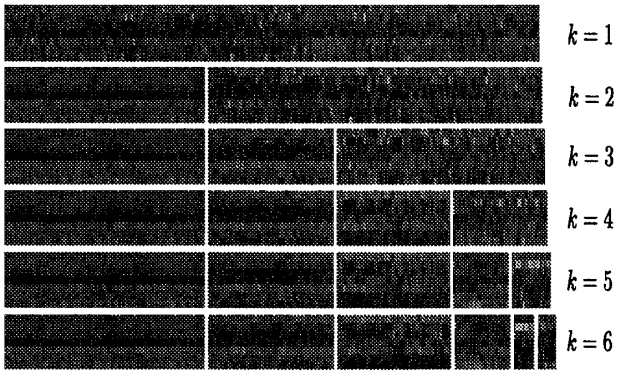

図 2: 可視化結果

6 まとめ

本研究では遺伝子発現データに対してスペクトラルクラスタリングを行い，それを階層的に可視化した．このことにより同クラスタ内のデータの特徴，そしてデータがクラスタリングされる様子を視覚的に捉えることができるようになった．また，本手法は k-means 法やファジィクラスタリングなど様々なクラスタリングにおいても対応でき，非常に柔軟な手法であると言える．

今後の課題として，他のクラスタリング手法についても考慮すること，任意のクラスタにおいても分割できるようにし，考察することが考えられる．

参考文献

[1] Michael B.Eisen et al., "Cluster analysis and display of genome-wide expression patterns," Genetics, Vol.95, pp.14863-14868, 1998.

[2] Cho,R,J. et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," Statistica Sinica, pp.241-262, 1998.

[3] Ulrike von Luxburg, "A tutorial on spectral clustering," Statistics and Computing, Vol.17, pp.395-416, 2007