3A1-01

包絡分析法を用いたクラスタリング手法の提案
－遺伝子発現データを対象として－

A Step Towards New Clustering Algorithm using DEA
－ for Gene Expression Data －

星埜　雅子[*1]　大野　博之[*2]　稲積　宏誠[*2]
Masako Hoshino　Hiroyuki Oono　Hiroshige Inazumi

[*1]青山学院大学大学院　理工学研究科　理工学専攻　知能情報コース
Graduate school of Science and Engineering, Aoyama Gakuin University

[*2]青山学院大学　理工学部　情報テクノロジー学科
Department of Science and Engineering, Aoyama Gakuin University

DNA microarray technology has now made it possible to monitor the expression levels of thousands of genes simultaneously. Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, there are factors which increase the challenges of comprehending and interpreting the resulting mass of data. A first step toward addressing this challenge is the use of clustering techniques. In this paper, we present a clustering algorithm using Data Envelopment Analysis (DEA). This clustering groups samples which has similar expression patterns for attribute genes.

1. はじめに

DNA マイクロアレイの発達により、何千もの遺伝子の発現値を一度に観測できるようになった。DNA マイクロアレイから得られた発現データを行列形式で表現した遺伝子発現データを用いて、複数の検体における遺伝子発現レベルの相違や、薬剤などの外部からの刺激による各遺伝子の発現レベルの変動を比較して観察することができる。

遺伝子発現データ解析の目的として、与えられたデータセットの分類が挙げられる。ここでの分類とは、ある特徴に基づいて遺伝子を分類する場合と、遺伝子を属性としてサンプル（検体）を分類する場合の 2 通りが考えられる。また、サンプル分類においては、既知のクラスに分類すると同時に、各クラスを細分化するサブクラスを発見することが必要とされている。特に、サブクラスの発見では一定の分類能力を有する遺伝子の抽出と同時に、有効なクラスタリング手法が必要となる。

本稿では、包絡分析法を用いた新しいクラスタリング手法を、遺伝子の発現傾向に基づく分類について検討する。

2. 包絡分析法 (DEA) の基本概念

包絡分析法 (DEA)([刀根 00]) は、任意の多入力多出力事例に対して多目的最適化問題を解くことにより、事例の効率性を相対評価する手法である。DEA では、より少ない入力で多くの出力を得る事例が効率的と判断する。通常経営分野では、事例は企業・学校などの組織体を指し、DMU と呼ばれる。例えば DMU_A の入力に対して得られる出力が他のどの DMU よりも大きい場合、DMU_A を「効率的」と評価し、同じ入力から DMU_A 以上の出力を持つ DMU がある場合、DMU_A を「非効率的」と評価する。非効率的な DMU に対しては、自らが効率的となるために参考にすべき効率的 DMU の集合を優位集合として与える。そして優位集合内の各 DMU に対して、参照度が算出される。

連絡先: 星埜 雅子，青山学院大学大学院
〒 229-8558 神奈川県相模原市淵野辺 5-10-1
E-mail: m-hoshino@ina-lab.it.aoyama.ac.jp

DMU_j $(j = 1, 2, \cdots, n)$ が、m 個の入力属性を表す入力ベクトル $x_{ij}(x_{1j}, x_{2j}, \cdots, x_{mj})$ と s 個の出力属性を表す出力ベクトル $y_{rj}(y_{1j}, y_{2j}, \cdots, y_{sj})$ をもつとき、DMU_k の効率は次式のように表せる。

$$
\begin{aligned}
&\text{目的関数} \quad Min\ \theta_k \\
&\text{制約式} \quad \theta_k x_{ik} - \Sigma_j(\lambda_j x_{ij}) \geq 0 \quad (i = 1, 2, \cdots, m) \\
&\qquad\qquad y_{rk} - \Sigma_j(\lambda_j y_{rj}) \leq 0 \quad (r = 1, 2, \cdots, s) \\
&\qquad\qquad \lambda_j \geq 0 \quad (j = 1, 2, \cdots, n)
\end{aligned}
$$

この線形計画問題を解いた結果得られる θ_k $(0 < \theta_k \leq 1)$ は、DMU_k の効率値を表し、$\theta_k = 1$ のとき DMU_k は効率的であり、$0 < \theta_k < 1$ のときは非効率的であることを示す。DMU_k が非効率的であるとき、λ_j は DMU_k から効率的な DMU_j への参照度を表す。したがって非効率的 DMU は、優位集合内の効率的 DMU と類似しているが相対的に弱い入出力関係を持っていると考えることができる。

3. DEA による遺伝子発現データの分析

3.1 DEA を用いた属性抽出法

遺伝子の発現傾向に基づくサンプルのサブクラス発見を行うために、遺伝子発現データのサンプルを DMU、マイクロアレイによって発現値を測定した遺伝子を属性として DEA を適用する。

既知クラスに対するサンプル分類能力の高い属性を抽出するためには、膨大な候補遺伝子のうち、既知クラスに分類する際の情報利得の大きな遺伝子を属性候補とすることが考えられる。その際、サンプルごとの発現値に大きなバラつきがあることが多い。図 1 は、後述する実験で用いたクラス分類能力の高い 2 つの遺伝子 (No.2642 と No.1902) が各サンプルにおいてどのように発現しているかを示した発現分布である。この図で示す 2 つの遺伝子は、発現値に差はあるが、似たような発現傾向を示す。この両方を属性として用いるのではなく、より大きく特徴を表している（発現している）遺伝子 No.2642 の

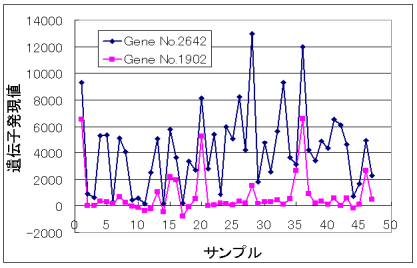

図 1: 2 つの遺伝子についての遺伝子発現分布

みを抽出することにより、異なる特徴を表す遺伝子を属性とする。その結果、明確に特徴を示し、互いに発現傾向が異なる属性を集めた属性群を用いてサンプル分類を実施できる。

類似の発現傾向を持つ遺伝子を見つけ、その中で相対的に高い発現値を示す遺伝子を探すことは、DEA における効率的事例を特定することと等価である。遺伝子発現データに対する DEA を用いた属性抽出は、以下の手順にしたがって行なう。

1. サンプルを既知クラスに分類する能力を情報利得により評価し、分類能力が上位の遺伝子を抜き出すことにより、遺伝子数を数千から数百に絞り込む。
2. 手順 1 において選んだ遺伝子を DMU、サンプルを属性として DEA を適用する。
3. 手順 2 の結果効率的な遺伝子の中で、一定以上の情報利得を有するものをサンプル分類で用いる属性とする。

例えば No.2642 は DEA により効率的と評価され、No.1902 は No.2642 を参照し非効率的と評価される。このように、優位集合から類似の発現傾向を持つ遺伝子を特定することにより、分類能力と表現能力の両方を加味した属性抽出が可能となる。

3.2 DEA を用いたクラスタリング手法

全遺伝子から抽出した遺伝子を属性、サンプルを DMU として DEA を適用する。その結果得られる参照関係に基づいて、発現傾向の類似しているサンプル同士を同じグループに分類するクラスタリングを実施する。なお、以下では、DEA が効率的と評価した DMU を「効率的サンプル」、非効率的と評価した DMU を「非効率的サンプル」と呼ぶ。

1. 効率値 α 以上の各非効率的サンプルは、しきい値 k 以上の参照度で参照している効率的サンプルと同じ性質を継承しているとし、この 2 つのサンプルを連結する。連結されたサンプル集合を一つのクラスタとする。
2. 手順 1 において、非効率的サンプルと対を作らない効率的サンプルを「孤立サンプル」と呼ぶ。孤立サンプル以外の効率的サンプルを対象サンプルから除去し、残ったサンプルを DMU として再び DEA を適用する。その結果に対して手順 1 を実行する。
3. 全てのサンプルがいずれかのクラスタに分類されるか、または DEA を適用した結果全ての DMU が効率的となるまで、手順 1〜2 を繰り返す。

手順 1 で効率値を考慮したのは、傾向の類似性と値のギャップの大きいものの判定を先送りすることを意味している。効率値、参照度のしきい値、手順 3 の結果残る孤立サンプルの取り扱いについては、様々な指標の導入が可能である。さらに、各段階でそれらを可変とするなどの取り扱いも可能である。

4. 実験

7129 個の遺伝子から成る急性リンパ性白血病 (ALL) と急性骨髄性白血病 (AML)72 検体に対する遺伝子発現データサンプル ([Golub 99]) に本手法を適用し、ALL のサブクラス発見を試みた。

2.1 節の遺伝子抽出方法と情報利得のみの 2 通りの方法によって抽出した 20 個の遺伝子を属性として、クラスタリングを行った。

実験データには、B-cell と T-cell という 2 種類の ALL サンプルが含まれていることが知られている。本手法により、2 種類の T-cell クラスタが発見され、他の ALL サンプルと異なる発現傾向があるということが示された。

B-cell サンプルのみから成るサブクラスと T-cell サンプルのみから成るサブクラスから、参照関係で結ばれたサンプルを 2 つづつ取り出し、属性として用いた各遺伝子に関する発現分布をそれぞれ図 2 と図 3 に示した。同じ ALL サンプルにも関わらず、2 つの図に示された発現パターンは大きく異なることが視覚的に分かる。また、図中の ALL44 と ALL9 は効率的サンプルである。そして、それぞれを参照している ALL12 や ALL11 は類似の発現パターンを示すが、その発現値は相対的に弱い。このように DEA を用いたクラスタリングでは、値に差はあっても、パターンの類似したサンプルを同じサブクラスに分類している。

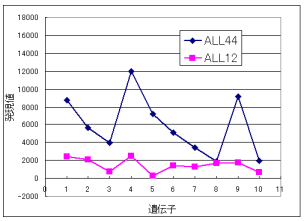

図 2: B-cell サンプル

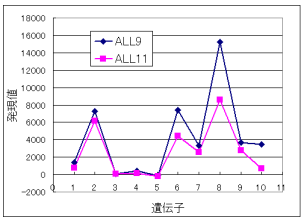

図 3: T-cell サンプル

5. 結論

本研究では、DEA を用いたクラスタリングを実施することにより、属性が示す特徴パターンの類似性による分類を行なった。実験により、このクラスタリングは 2 種類の ALL サンプルの特徴を分類することができた。今後は、効率値や参照度のしきい値の決め方、孤立サンプルの取り扱いについて検討する。

参考文献

[刀根 00] 刀根 薫, 上田 徹: 経営効率評価ハンドブック -包絡分析法の理論と応用-, 朝倉書店 (2000).

[Golub 99] Golub, T. R. et.al.: Molecular classification of cancer -Class discovery and class prediction by gene expression monitoring-, pp.531-537, Science 286 (1999).