包絡分析法を用いた遺伝子発現データ解析の試み

星埜　雅子†　　稲積　宏誠††
青山学院大学理工学部

１．はじめに

2001 年 2 月に国際解析チームとセレラ社からそれぞれヒトゲノム配列の概要が報告され、60 以上の生物種のゲノム配列が決定された。その結果、遺伝子発現解析、タンパク質の構造決定等をはじめとする、いわゆるポストゲノムシークエンス研究が本格化してきた。これらの研究においては、膨大で多種多様な生物情報を効率よく整理・解析し、その生物学・医学的意味を明らかにする事が必要であり、バイオインフォマティクスが必要不可欠となっている。

バイオインフォマティクスは、生命科学に情報科学的な視点や概念を導入した研究分野であり、「遺伝子・タンパク質の構造解析」、「遺伝子・タンパク質の相互作用の解析」、「細胞・組織・固体レベルの解析」に分類できる。遺伝子・タンパク質の相互作用、すなわち、関係性の研究におけるアプローチとして、環境変化によるゲノム変異によりもたらされる細胞システムの変化を遺伝子発現プロフィールにより説明付けるというものがある。これにより、機能未知の遺伝子の機能を予測することができ、さまざまな応用分野への応用が期待されている。

本研究では、遺伝子発現データを用いた疾患分類の問題に注目し、分類能力の高い遺伝子の抽出法とそれを用いた未知検体（細胞）の分類のための新たな手法を提案する。特に、分類方法については、主として経営分野で多目的最適化問題に利用されている包絡分析法（DEA）に注目し、通常のクラスタリング手法と組み合わせたハイブリッド分類となっている。

２．遺伝子抽出

遺伝子発現データは、実験等の制約により事例数に対して属性数が膨大となるため、観測された全ての遺伝子から分類能力の高い属性のみを抽出する必要がある。たとえば、疾患分類を行う際には、異なる疾患をもつ事例を正確に分類できる発現パターンをもつ遺伝子を抽出することが要求される。このような手法についてはすでにいくつか提案されており[1][2]、抽出された遺伝子を用いた際の誤分類率を用いてその手法の優劣が評価されている。本研究では、C5.0 で用いられている情報量に基づいた基準を用いて、分類能力の高い順に選択する方法を用いた。なお本稿では、遺伝子抽出法についての評価、検討は省略する。

３．DEA による発現データの分類

DEA とは、任意の多入力多出力事例の多目的最適化問題を解析するための一手法である[3]。DEA ではまず各事例は自らが最も入出力比率が効率的となるような各入力･出力に対するウェイトを設定する。次にそれを用いて他の事業体を評価し、全ての事業体の中で、より効率的なものが存在しない場合には、自らの活動を効率的と判断するものとし、この条件を満たす事例を「効率的である」とする。したがって、全ての事例は効率的である事例とある効率的事例よりも劣ることが示されている事例のいずれかとなる。その結果、各非効率的事例には、優位集合と呼ばれる目指すべき効率的な事例が対応づけられることになり、効率的事例を中心にして全事例をその入出力関係の特徴による(重複を許す)グループに分類することができる。

DEA には多くのモデルが存在するが、最も一般的な CCR モデルを用いると、各事例の入力データベクトル $\boldsymbol{x}_j$、出力データベクトルを $\boldsymbol{y}_j$、全ての事業体を示す事例集合を $\mathbf{X}=(\boldsymbol{x}_j)$、$\mathbf{Y}=(\boldsymbol{y}_j)$として次式のように表すことができる。ただし、j= o を分析対象とする。

＜LPo＞　目的関数　：Min θ

制約式：$\theta \boldsymbol{x}_o - \mathbf{X}\boldsymbol{\lambda} \geqq 0$

$\boldsymbol{y}_o - \mathbf{Y}\boldsymbol{\lambda} \geqq 0$

$\boldsymbol{\lambda} \geqq 0$

これを全ての事例について解くことによって、θが事例 o の効率値(効率的である場合には１となる)、$\boldsymbol{\lambda}=(\lambda_j)$は o が非効率なときに他の事例効率的な事例を参照する度合いを表す。

これを遺伝子発現データサンプルの疾患分類に適用するために、遺伝子発現データを「選択された遺伝子発現値の逆数を入力とし、出力を全て等しい事例」と解釈する。これに対して DEA を適用すると、疾患がある小さな入力（大きな発現値)の組合せについてある特徴をもつグループに分類されることになる。その結果、その疾患を特徴づける遺伝子の組み合わせ方の特徴によるグループにより分類することができると考え

A Step towards gene expression data analysis using DEA
†Masako Hoshino Aoyama Gakuin University
‡†Hiroshige Inazumi Aoyama Gakuin University

られる。

ここでは、$\max_j, \min_j$ はそれぞれ遺伝子 j の最大、最小発現値として以下の式によりサンプル i の遺伝子 j の発現値 x_{ij} を変換する。

$$z_{ij} = 2 - (x_{ij} - \min_j) / (\max_j - \min_j)$$

その結果 $z_{ij} \in [1, 2]$ となり、出力値 = 1 とすることにより DEA を適用可能となる。

DEA によって得られた参照関係にもとづいて次のようにグループ分けを行う。

①非効率的であるサンプルにおいてしきい値 k 以上の参照度を持つ効率的サンプルが存在するならば、それらを同じ性質を継承する対とし、全てのサンプルにおいて連結された複数のサンプル集合をグループとする。

②①の結果グループに属さない非効率的サンプルのなかで、グループに属する効率的サンプルへの参照度が上位 m 個以内に存在するものがあれば、そのサンプルはその効率的サンプルの属するグループに含める。

③②において上位 m 個以内にグループに属する効率的サンプルがなければ、ここで選択された効率的サンプルを中心として①と同様にしてグループを作成する。

④①-③で複数のサンプルで連結されないものを孤立データとする。

４．ハイブリッド分類法

一般的に行われている発現データサンプルの分類ではクラスタリングが多く用いられている。これは、各発現パターンの距離尺度による類似性による分類とみなすことができる。一方 DEA の参照関係によるグループ分けは、距離尺度ではなく発現パターンの特徴の類似性による分類とみなすことができる。クラスタリング結果とグループは全く異なった分類結果を示す。そこで、未知の発現データサンプルに対する分類方法として次のようなアルゴリズムを提案する。

① 訓練データによりつくられたクラスタリングにより未知発現データの疾患を決定する。

② 訓練データと未知データを用いて DEA によるグループ化を行った結果、未知サンプルの属するグループが①で得られた疾患と異なる疾患を示していたならば保留と判断する。

以上の結果、ある分類器による分類結果の中に保留を設定することによって誤分類を防ぐのがハイブリッド法のねらいである。

５．実験

実験データとしては、7129 遺伝子から成る急性骨髄性白血病（AML）と急性リンパ性白血病（ALL）73 検体の発現データサンプルを用いる。まず 20 個の遺伝子を抽出した後にハイブリッド分類を適用した分割交差検定結果を以下に示す。

表１　ハイブリッド分類結果例

	1	2	3	4	5
A	*L3*, L4, *L5*, *L9*, *L31*, *L32*, *L42*, *L47*			L11, *L12*, *L20*, *L22*, **L23**, *L24*, *L25*, *L26*, L27, *L30*, *L41*, *L46*	**L13**, **L38**, **L43**
B	L16, **L18**, *L40*				
C	*L39*				**L44**
D	*L2*, *L15*, L17, *L19*, **L34**				
E	*L35*, *L36*				**L7**
F		**M8**		*L45*	
G		**M3**, **M23**	*M1*, *M2*, *M4*, *M5*, *M6*, *M7*, *M9*, *M10*, *M11*, **M13**, *M14*, M17, *M18*, *M19*, *M21*	*L6*, M22	
孤立	**L1**, **L14**, **L28**, **L29**	**M12**	**L10**, **M15**, **M16**, **M20**, **M24**	**L21**	**L8**, **L37**, **L48**

この表は、行方向に DEA グループ、列方向にクラスタの内訳を示す。尚、ALL を L、AML を M とし、その後ろにサンプル番号を付与する事で各サンプル名を簡略化した。下線はテストサンプル、太字は効率的サンプル、斜体は非効率的サンプルをそれぞれ表す。

この結果、保留という概念を導入した事で、クラスタリングにおいて平均 1.9 個であった誤分類サンプル数が、平均１個に減少した。

６．結論

本稿では、DEA を遺伝子発現データに適用する方法を示し、ハイブリッド分類法により、疾患分類において誤分類を抑制することができることを示した。今後は、本手法の理論的な位置づけの明確化と従来の分類法では発見することが困難な問題への適用可能性について検討する。

参考文献

[1] Piatetsky-Shapiro,T. et.al. "Capturing Best Practice for Microarray Gene Expression Data Analysis", *Proc. of KDD-2003*, pp. 407-415, (2003).

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

[3] 刀根薫，上田徹 監訳 :「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，pp. 1-75, 371-380, (2000).