判別分析を用いた形質への影響たんぱく質を発見するデータマイニング手法

横山恵樹[††] 井上悦子[†] 吉廣卓哉[†] 中川優[†]

[†] 和歌山大学大学院システム工学研究科

[††] 和歌山大学システム工学部

1. はじめに

本研究では、二次元電気泳動により得られたたんぱく質発現量データから、特定の形質に影響を与えるたんぱく質を推定するデータマイニング手法を提案する。

通常たんぱく質は複数個が影響しあって形質を決定付けると考えられている。本手法では、二つのたんぱく質発現量を軸とする二次元空間上で特定の形質を線形分離できるような二つのたんぱく質を、判別分析を用いて網羅的に探索することで、その形質に何らかの影響を及ぼすと推定されるたんぱく質を発見する。判別分析は、はずれ値を含むデータであっても統計的に分離度合いを数値化できるため、ノイズを含み得るデータを網羅的に探索するのに適している。また、分析結果の閲覧ツールを併せて開発することで、値だけでは分からない判別結果の傾向確認が可能になった。

2. たんぱく質発現量データと形質データ

本研究で用いるたんぱく質発現定量データは、試料（サンプル）を二次元電気泳動することによりたんぱく質を分離し、分離した各スポットの容積（これはスポットの面積や濃淡から計算できる）を画像解析により計算した結果を用いる。比較する全サンプルの電気泳動画像に対してこれを計算した後、同じたんぱく質に対応するスポットを各画像から探しマッチさせる。これにより、全サンプルに対して、そこに含まれる各たんぱく質の発現量が求まり、この発現定量データを用いてデータマイニングを行う。つまり、発現定量データとはサンプルとたんぱく質の二次元の表形式で表され、表の各セルには、対応するサンプルにおける各たんぱく質の発現量が格納される。但し、各スポットの体積は画像の濃淡の影響を受け個体差が出るため、各スポットの発現量をスポット全体の発現量の総和に対する割合とするなどの正規化が必要である。

形質データは実験対象の各個体を測定して得られる値であり、個体と測定項目の二次元の表形式で与えられる。ここで個体は発現定量データのサンプルと対応していることとする。

本研究で扱う形質データは、例えば植物の背丈や茎径のような連続値データと、牛肉の脂肪交雑度合（BMS: 複数段階の評価結果として得られる）のような離散値データに分類される。遺伝子型のような値の大小がないカテゴリデータは本研究では対象外とする。本研究で想定するデータは、たんぱく質数が数百程度、サンプル数が数十以上のたんぱく質発現量データとする。

3. 提案するデータマイニング手法

提案手法は二次元、二群の線形判別分析[2]を基にしている。形質が連続値や離散値である場合には、そのままでは判別分析を適用することができないため、形質に閾値を設けることでサンプルを２群に分割する。基本的な処理の流れは次の通りである。全ての２たんぱく質の組合せに対して、いくつかの閾値でサンプルを２群に分割した後に判別分析を適用する。その結果得られる適合率（２群がどれだけきれいに線形分離されたかを表す）を用いて組合せをランキング化し、上位のものを取り出すことで、その形質に深く関係するたんぱく質の組合せを取り出す。

形質に対する閾値の設定は、サンプルを形質値でソートした後、サンプル数が均等になるように十分割する各境界値を用いることとした。判別分析により得られる適合率は、閾値が少し変化した場合には大きく変化することはない。このため、適当な粒度でいくつかの閾値を設定する方法で十分である。

形質に影響するたんぱく質の機能の探求においては、ある程度少数のたんぱく質の発現量が形質と関係している場合を発見することが肝要

Data Mining Technique using Discriminant Analysis to Find Proteins that Influence Phenotypes

Keiju YOKOYAMA[†], Etsuko INOUE[††],
Takuya YOSHIHIRO[††], Masaru NAKAGAWA[††]

[†] Graduate School of Systems Engineering, Wakayama University

[††] Faculty of Systems Engineering, Wakayama University

である。たんぱく質数が少数であれば、その後の *in vitro* な実験によりたんぱく質の機能を解明することが容易になる。本手法において、２つのたんぱく質の組合せに対して適合率が高いことは、これらのたんぱく質が何らかの形でその形質に影響を与えている可能性を示唆している。

しかし、適合率が高いというだけでは、実際のデータマイニング作業においては、そのたんぱく質の組合せが真に興味深い関係を表しているかを判断することはできない。実際に散布図を表示し、はずれ値やサンプルの分布を見てはじめて、今後探求を進めるべき興味深い関係であることが判断できる。このため、実用的に利用できるマイニングインタフェースを併せて実装した。実装する主なインタフェースは、形質ごとに用意されたたんぱく質の組合せのランキング画面と、選ばれた二つたんぱく質を軸に取った散布図画面である。散布図画面からは、サンプルを２群に分割した閾値を少しずらした場合の結果などの、いわば近傍結果も確認できるようにすることで、データの傾向をインタラクティブに閲覧できるように工夫してある。

4. 実装

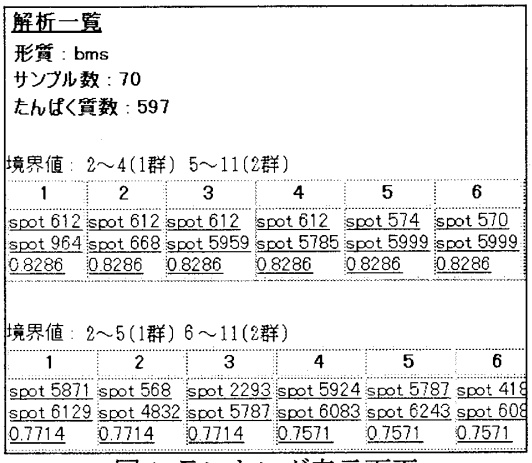
本システムは Web アプリケーションとして試作した。判別分析には統計ライブラリである R[3]を用い、スクリプト言語は PHP を使用した。

図 1 は選択された形質のランキング表示したものである。ランキングは形質の閾値毎に作成され、各セルにはたんぱく質の組み合わせとその適合率が表示される。各セルは詳細画面（散布図）へのリンクになっている。

図 2 は詳細画面である。ランキング表示画面で選択されたたんぱく質の組み合わせを二軸に取り、色分けにより元データがどちらの群に属していたか知ることができる。またグラデーション表示することにより二群に分ける前の元データの形質情報も併せて閲覧できる。散布図中に引かれている直線は判別分析により二群に分けた際の判別直線である。また、画面右下には形質データから判断された二群と判別結果の正誤表を表示し、判別結果を確認できる。

5. 結果

評価として、ウシの発現量解析結果を適用した。形質は 12 段階の離散値データである脂肪交雑度(BMS)、たんぱく質数は約 800、サンプル数は 70 である。その結果、例えば図２のような、単独のたんぱく質だけでは説明のつかない形質への影響を複数発見することができた。



解析一覧

形質：bms
サンプル数：70
たんぱく質数：597

境界値： 2～4(1群) 5～11(2群)

1	2	3	4	5	6
spot 612	spot 612	spot 612	spot 612	spot 574	spot 570
spot 964	spot 668	spot 5959	spot 5785	spot 5999	spot 5999
0.8286	0.8286	0.8286	0.8286	0.8286	0.8286

境界値： 2～5(1群) 6～11(2群)

1	2	3	4	5	6
spot 5871	spot 568	spot 2293	spot 5924	spot 5787	spot 418
spot 6129	spot 4832	spot 5787	spot 6083	spot 6243	spot 608
0.7714	0.7714	0.7714	0.7571	0.7571	0.7571

図 1. ランキング表示画面

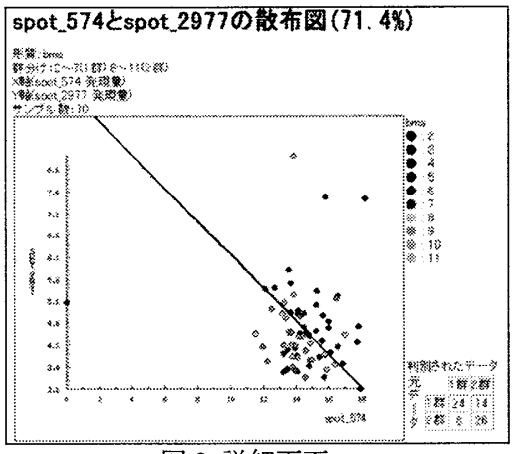


図 2. 詳細画面

6. おわりに

二群での線形判別分析を用いて形質に影響するたんぱく質を発見するデータマイニング手法を提案し、実用的なツールとして実装した。発現量データを線形分離する手法として、サポートベクターマシン[4]がある。しかし、今回想定するような、ノイズ成分が多いデータの場合は、判別分析の方が適すると考えている。今後は実データに対しての適用をさらに進めることで、本手法の有効性を評価する予定である。

なお、本研究は和歌山県地域結集型共同研究事業により実施した。

参考文献

[1] 和歌山県地域結集型共同研究事業, http://www.wakayama-kessyu.com/ .
[2] 長谷川勝也, “ゼロからはじめてよくわかる多変量解析,” 技術評論者, 2004.
[3] 舟尾暢男, “The R Tips,” 九天社, 2005.
[4] 元田浩, 津本周作, 山口高平, 沼尾正行, “データマイニングの基礎,” オーム社, 2006.