3J-1

発現定量データを用いた 判別分析によるデータマイニング支援システム

谷 正浩† 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

和歌山県地域結集型共同研究事業[1]では，ハイスループットなたんぱく質解析拠点を整備し，解析手法の一つとして二次元電気泳動画像を用いた大規模なたんぱく質の発現定量解析を実施している．本研究はその中で、発現定量データと形質データを用いた高次元の判別分析[2]を行うソフトウェアの試作を行った。たんぱく質研究においては、単に高次元の判別分析結果を数値として得るだけでなく、分布を確認するために可視化できる必要がある。そこで本研究では、例えば三次元のデータを用いて分析する場合には、三次元目のたんぱく質の発現量がある一定の連続範囲にあるサンプルのみを用いて二次元の判別分析をすることとし、この二次元の判別結果が良くなるような三次元目の発現量の範囲を見つけて二次元の散布図を用いて可視化する。本稿では、このような擬似的な高次元判別分析を行う手法を提案し、これを試作したので報告する。

2. 発現定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。比較する全てのサンプル（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。解析したい全てのスポットを各画像から探してマッチさせることで、分析に用いる発現定量データが得られる。つまり、発現定量データとはサンプルとたんぱく質の 2 次元の表形式で表され、表の各セルには、対応するサンプルの電気泳動画像における各スポット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。また、本研究では判別分析を行うので、各セルに格納される測定値はカテゴリ値であり、カテゴリを表すいくつかの値の中の一つをとることとする。

3. データマイニング手法

提案手法は二次元の判別分析を基にしている。二つの次元にそれぞれ二つのたんぱく質の発現量を対応させる。サンプルを複数の群に分けるために形質データを用いる。判別線により群がきれいに分けられるほど適合率が高くなる。ここにもう一次元を加えて、三次元目のたんぱく質の発現量が一定の範囲のサンプルのみに限定することで適合率が上がるような場合を発見することが目的である。これは、例えば三次元目のたんぱく質が多く発現すると一、二次元目のたんぱく質や形質に影響する場合に、その影響を除いた分析結果を得られるという効果がある。

しかし，この手法で解析を行った場合に問題点がある．例えば，三次元目がある範囲の時に適合率が良いとすると，そこから少しずれた範囲の解析結果もほぼ等しくなる．また，適合率が高いが含まれる個体が比較的少ない範囲と，適合率はそれより少し低いが多くの個体を含む範囲を比較した場合，必ずしも適合率が良いものがいい結果だとは言い切れない。全ての範囲に対して適合率を計算した後、ユーザにどのように提示するかを工夫する必要がある。

そこで本研究では、三軸に対する全てのたんぱく質の組合せのそれぞれに対して適合率が最適となる範囲を求めておき、解析したい形質に

A System to Support Data Mining based on Discriminant Analysis from Protein Quantitative Data

Masahiro TANI† Etsuko INOUE†
Takuya YOSHIHIRO‡ Masaru NAKAGAWA‡
† Graduate School of Systems Engineering, Wakayama University
‡ Faculty of Systems Engineering, Wakayama University

対してその値が高い順にランキング表示する。つまりユーザは、形質を選択すれば適合率の高い順に三軸の組合せを知ることができる。そして、表示された組合せのいずれかを選択することで、その組合せの各範囲がどのような適合率になっているかを概観できる。さらに、範囲を一つ選択することで判別結果を二次元の散布図上で確認することができる。このインタフェースを実装することにより、適合率の高い組合せを効率良く発見でき、また興味に応じて散布図まで詳しく閲覧することができる。

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。スクリプト言語の PHP を使用した.

ユーザは，解析に使用するたんぱく質の発現定量データと形質データをファイル登録画面からサーバにアップロードする．サーバでは，アップロードされたデータを用いて解析を行う．ユーザは解析結果を「4.1 ランキング表表示画面」と「4.2 詳細表表示画面」で閲覧することができる.

4.1 ランキング表表示画面

ランキング表表示画面では，形質毎に、三軸の全組合せに対して最適適合率を計算し，これをランキング表示している.

ランキング表の例を図 1 に示す。この例は形質としてウシの肉質を用いた例であるが、「きめ」「BMS」「BFS」の各形質項目に対して、3 位までの順位が表示されている（実際は指定した順位までの表示が可能である）。この例では、「きめ」に対しては一、二次元目にたんぱく質 1 と 10、三次元目にはたんぱく質 3 を選んだ場合の適合率が最も高く 82.5%であり、その時の三次元目の範囲は 0.33 から 22.67 までであることがわかる。ランキングの各セルはリンクになっており、クリックすることで詳細表表示画面（4.2 節）に遷移する。

	1位	2位	3位
きめ	第1,2軸:protein1,10 範囲:0.33<protein3<22.67 適合率:82.5%	第1,2軸:protein5,30 範囲:0.67<protein5<9.67 適合率:76.5%	第1,2軸:protein5,27 範囲:0.82<protein13<20 適合率:75.6%
BMS	第1,2軸:protein8,12 範囲:2.12<protein9<12.5 適合率:91.2%	第1,2軸:protein12,17 範囲:0.51<protein15<19.67 適合率:90.8%	第1,2軸:protein2,3 範囲:0.33<protein1<20 適合率:88.6%
BFS	第1,2軸:protein12,13 範囲:4.1<protein8<22.83 適合率:95.4%	第1,2軸:protein15,21 範囲:1.67<protein26<15.64 適合率:94.1%	第1,2軸:protein12,28 範囲:3.2<protein3<18.4 適合率:89.6%

図 1 . ランキング表の例

4.2 詳細表表示画面

詳細表表示画面ではランキング表表示画面（4.1 節）で選択された三軸の組合せに対して、三次元目の範囲全てについてのスコアを表形式で表示している．ただし範囲の組合せが膨大なので、ランキング表表示画面とは異なりスコアや範囲は表示させず，スコアに応じて各セルに色をつけている．色は，スコアが良いものは濃く，悪いものは淡くしている.

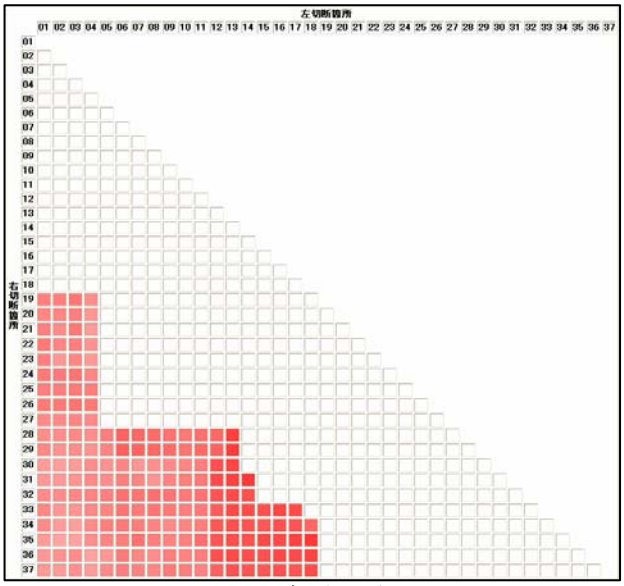

図 2. 詳細表の例

詳細表の例を図 2 に示す．表の行と列は，それぞれ三軸目の範囲の始点と終点を示している。数値はサンプルの個数だけ振られるが、これは三軸目の値でサンプルをソートした時にいくつめのサンプルの間を境界とするかを示している。つまり、各セルは対応する数値番目のサンプル間を始点・終点とした範囲の結果を表している。図 2 は 40 サンプル程度の例であるが、色の濃い部分が近くに集まっていることがわかる。

5. まとめ

判別分析を用いて可視化可能な三次元の判別分析を行うデータマイニング支援システムを設計・構築した．今後、実データを用いて分析を行うことで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で実施した。

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] 奥野忠一，“多変量解析法<改訂版>,” 日科技連, 2005.