1998年度日本オペレーションズ・リサーチ学会
秋季研究発表会

整数計画法による最適線形判別関数

01202720 成蹊大学 新村秀一 SHINMURA Shuichi

1 はじめに

判別分析は、重回帰分析と並んで応用範囲の広い手法である。今後は、データベースマッケーティングやデータマイニングなどの目的で集められた大量のデータと変数に対して、適切な判別関数を作成することが重要になってくる。

この目的のため、数理計画法を用いた新しい判別関数を提案し（LP判別関数、IP判別関数）、従来のFisherの線形判別関数（F判別関数）と比較する。これによって、従来の判別分析では分からなかった新しい知見が得られた。

2 方法

2・1 Fisherの線形判別関数

F線形判別関数は、p個の説明変数をもつ2群が、p変量正規分布に従う確率密度関数をもつとする。2群の分散共分散行列が同じ（$\Sigma_1=\Sigma_2=\Sigma$）として、次の尤度比 $f_1(\mathbf{x})/f_2(\mathbf{x})$ の対数をとったものを考える。

$$F(\mathbf{x})=\log[f_1(\mathbf{x})/f_2(\mathbf{x})]=\{\mathbf{x}-(\mathbf{m}_1+\mathbf{m}_2)/2\}'\Sigma^{-1}(\mathbf{m}_1-\mathbf{m}_2)$$

このF(x)は、$\mathbf{x}'\beta+\beta_0$というxの線形関数になっており、これをFisherの線形判別関数と呼ぶ。この時、尤度比が1すなわち2群の確率密度が等しい点を尤度比方式による判別境界点と呼び、F(x)＝0は、p次元のデータ空間を2分する超平面になる。

一方、2群のデータ件数をn_1とn_2として、データを1群と2群の順に並べ替えて、新しい目的変数yの値として1群に$1/n_1$、2群に$-1/n_2$を与える。この時、回帰分析によって得られる回帰係数は、Fisherの線形判別関数と比例関係にある。すなわち、判別分析は回帰分析の特殊応用例に還元される。

Fisherの線形判別関数を、$f(\mathbf{x})=a_1x_1+a_2x_2+\cdots+a_px_p+a_0$ とし、1群に属するデータ$\mathbf{x}$をこの式に代入し、$f_1(\mathbf{x})=a_1x_1+a_2x_2+\cdots+a_px_p+a_0\geqq 0$ であれば、正しく1群に判別されたとする。$f_1(\mathbf{x})<0$ であれば、2群に間違って判別（誤分類）されたものとする。関数の添え字の1は、1群のデータを用いていることを示す。

2群に属するデータ$\mathbf{x}$の場合は、$f_2(\mathbf{x})<0$で正しく分類され、$f_2(\mathbf{x})\geqq 0$で誤分類されることになる。

三宅・新村（文献1-3）は、定数項を1に規準化した $f(\mathbf{x})=a_1x_1+a_2x_2+\cdots+a_px_p+1$ をもちいて、個々のデータ$\mathbf{x}_i$から得られる線形式$\mathbf{Hi}$（$\mathbf{x}_i\mathbf{a}+1=0$）を導入した。各$\mathbf{Hi}$は判別係数の空間$\mathbf{a}$を2分する超平面である。この$\mathbf{Hi}$から作られる誤分類数を最小にする凸体を、総当たり法で探すアルゴリズムを提案したが、計算時間がかかり実用例への適用にいたらなかった。

2・2 最適線形判別関数

数理計画法でもって、回帰分析が行えることは広く知られた事実である（文献4－5）。

数理計画法を用いた判別関数は、LINDOの形式に従えば、次のように定義できる。最初のm個の制約式は1群のデータであり、その後のn個の制約式は2群のデータに対応している。ただし、左辺には定数項は許されないので、1は右辺に移項する必要がある。また、右辺には変数は許されないので、ce_iは左辺に移項する必要がある。

IP判別関数は、大きな定数cとe_iをINTコマンドで0/1型の整数という制限を加える。LP判別関数は、Cを1とし、e_iは実数とする。

```
MIN  e1+e2+…+em+e(m+1)+…+e(m+n)
ST
x1 ja1+…+xp ja p+1 >=ce i        e i  ∈  1群
−x1 ja1−…−xp ja p−1 >=ce j      e j  ∈  2群
END
```

LP判別関数は、誤分類されるケースの判別境界点からの距離の和を最小化している。判別境界点から離れたケースほど大きなウエイトを占める。IP判別関数は、誤分類されるケースの個数の和を最小化している。もしデータが通常のF判別関数が期待している理論的な仮定を満たしているなら、F判別関数は誤分類数を最小化するので、IP判別関数と一致するはずである。

3 結果

3・1 アイリスデータ

判別分析やクラスター分析の評価に良く用いられるフィッシャーのアイリスデータ(文献6)のバーシクルとバージニカの100個のデータを用いる。すなわち、2群を4個の説明変数で判別する問題である。

表1に、逐次変数選択法で得られた増加法(減少法と一致)によるモデル系列（上昇・下降基本系列という）での各判別関数による誤分類数をまとめてある。IP判別関数は、F判別関数よりも判別成績は良い。LP判別関数はIP判別関数と同じ事もあるが、F判別関数よりも悪くなることもある。

表1 アイリスの両基本系列での誤分類数

	1	2	3	4
IP	5	3	2	1
LP	5	7	2	1
F	6	5	4	3

3・2 医学データ

次に、児頭骨盤不均衡（Cephalo-Pelvic-Disproportion、略してCPD）という医学データを用いる。このデータは、日本医科大学第1病院の鈴村・松木・武井らによって集められた。180例の自然分娩群と60症例の帝王切開群のいずれの方法を選択するかを、19個の説明変数から事前に予測するために集められた(文献7)。

上昇・下降基本系列の誤分類数は、

- IP判別関数は、F判別関数よりも判別成績は良い。
- LP判別関数は、IP判別関数よりも悪い。一方、8変数（上昇）あるいは10変数(下降)まではF判別関数よりも悪くなることも良くなることもあるが、それ以上ではF判別関数よりも誤分類数は少なくなる。

表2は、19変数と、多重共線性のある3変数を省いた(文献8・9)16変数の上昇・下降基本系列での誤分類数をまとめたものである。これから、19変数の下降基本系列の成績が悪く、16変数の下降基本系列の成績が良いことが分かる。残りの2つは、その中間にある。しかし、ここで注意すべきは、16変数の下降基本系列の決定係数による順位は決してよくない点である。10変数から16変数の間では、19変数の下降基本系列は順位が1位であるのに対し、16変数のそれは20位よりも悪い。すなわち、**決定係数が良いことが誤分類数が少ないことに必ずしも対応していないという重大な知見が得られた。**

表2 各系列でのIP判別分析の誤分類数比較

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
19F	19	13	12	10	10	8	7	6	4	4	4	4	3	3	3	3	2	2	2
19B	19	13	12	11	11	9	9	8	6	6	4	4	3	3	3	3	2	2	2
16f	19	13	12	10	10	8	7	6	4	3	3	3	3	2	2	2			
16b	19	13	12	10	8	7	7	6	4	3	3	3	3	2	2	2			

4 まとめと今後の課題

今回、数理計画法を用いたIP判別関数とLP判別関数を提案した。

説明変数が少なく、比較的Fisherの線形判別関数の理論的前提を満たしていると思われるアイリスデータでは、それほど優位性は認められなかった。しかし、医学データでは、19個と説明変数の数が多く、そのため多重共線性があり、しかも正規分布でなかったり、離散変量であったりと種々のデータタイプが混在している。このようなデータでは、**IP判別関数はFisherの線形判別関数に対して、優位な成績が得られた。さらに、決定係数が悪くても誤分類数が少ないモデル系列が得られた。**

文献

1 三宅章彦、新村秀一：最適線形判別関数のアルゴリズムとその応用、医用電子と生体工学、18-1、15/20（1980）

2 A. Miyake& S. Shinmura : An algorithm for the optimal linear discriminant functions, Proceedings of the International Conference on Cybernetics and Society,1447/1450(1987)

3 S.Shinmura &A.Miyake : Optimal linear discriminant functions and their application, Proceedings of the COMPSAC'79, 167/172(1979)

4 L. Schrage（新村秀一、高森寛訳）：実践数理計画法—LINDOを用いて—、朝倉書店(1992)

5 新村秀一：LINDOを用いた線形回帰分析例、日本オペレーションズ・リサーチ学会秋季研究アブストラクト集、13/14(1984)

6 新村秀一：パソコン楽々統計学、講談社(1997)

7 松木玄篤：数理的解析によるC.P.D.の判定、日本産科婦人科学会雑誌、30-12、1727/1736(1978)

8 新村秀一、三宅章彦：重回帰分析と判別分析のモデル決定（1）—19変数をもつC.P.D.データの多重共線性の解消—、医療情報学3-3、107/123(1983)

9 新村秀一：重回帰分析と判別分析のモデル決定（2）—19変数をもつC.P.D.データのモデル決定—、成蹊大学経済学部論集27-1、180/203(1996)