1999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

整数計画法を用いた最適線形判別関数(OLDF)

1202720 成蹊大学 新村秀一 SHINMURA Shuichi

1 はじめに

整数計画法を用いて、内部標本の誤分類数を最小化する最適線形判別関数(IP-OLDF)を開発し、報告してきた(文献1-2)。内部標本の誤分類数は、母集団に比べ、説明変数が多くなるほど、あるいは標本数が少ないほど、過小評価される(文献4)。従来の線形判別関数や2次判別関数は、多次元正規分布を仮定しているが、現実のデータは、この仮定を満たすことは稀である。また、誤分類数は、事前確率やリスクの導入によって、同じデータであっても異なってくる。その点、IP-OLDFは、特定の理論分布を前提にせず、誤分類数が一意に決まるという利点がある。さらに説明変数の増加に対し、単調減少性を示す。

本手法を用いれば、分析対象のデータの多次元正規性からの乖離度や、多重共線性の問題点を視覚的に理解できるのではないかと期待できる。

また、データに対し一意に決まる誤分類数を用いて、従来の手法の評価が行える。

2 アルゴリズム

線形計画法を用いた判別関数は、回帰分析と同様、いくつか提案されている(文献5)。IP-OLDFを少し変形すれば、誤分類されたケースの判別境界点からの距離の和を最小化する、LP-OLDFが定式化できる。

IP-OLDFは、整数計画法(IP) は計算時間がかかると思われていたので、定式化はないようだ？

MIN $e_1+e_2+\cdots+e_m+e_{(m+1)}+\cdots+e_{(m+n)}$

ST

$x_{11}a_1+\ldots+x_{p1}a_p+1>-ce_1$

.

.

$x_{1m}a_1+\ldots+x_{pm}a_p+1>-ce_m$

$-x_{1(m+1)}a_1-\cdots-x_{p(m+1)}a_p-1>-ce_{(m+1)}$

.

.

$-x_{1(m+n)}a_1-\cdots-x_{p(m+n)}a_p-1>-ce_{(m+n)}$

END

e_i は 0/1 の整数変数。cは大きな正の実数。a_i は判別係数で、x_{ij} は変数 x_i のj番目のデータ。1 群は最初の m 件であり, 2 群はその後のn件である。

3 データ

これまで3種類のデータで検討を行っている。

Fisherのアイリスデータは、判別やクラスター分析の研究に良く取り上げられるが、変数が4個であり実際の評価には使えない事がわかった。Fisherの線形判別関数では誤分類数は3であるが、IP-OLDFは1である。 乱数データの成果は、別途報告する。

本稿では、医学データ(CPD、240人*19変数)の成果を報告する。このデータは、高度な多重共線性があり、3変数を省くと多重共線性が解消される(文献6)。

表1 AIC最小モデルにおける判別係数

	X9	X12	X15	X18	定数項
IP	-5.363E-2	-6.177E-3	1.341E-2	-1.249E-2	1
LP	-3.228E-3	-6.791E-4	8.071E-4	-3.823E-3	1
F(P)	-9.349E-3	-1.539E-3	1.602E-3	-3.740E-3	1
F(0.5)	-8.592E-3	-1.414E-3	1.472E-3	-3.437E-3	1

4 評価方法

IP-OLDFを、LP-OLDFや線形判別関数の誤分類数と基本系列上で比較した。IP-OLDFの誤分類数は、説明変数の増加と共に単調減少するのに対し、LP-OLDFは最初振動し、その後減少傾向を示した。線形判別関数は、最初振動し、それ以上では減少しなかった。

次に、19変数の上昇基本系列(19F)、減少基本系列(19B)と多重共線性を解消した16変数の基本系列(16f、16b)上で、IP-OLDFの誤分類数を比較した。この結果、16bのモデル系列は、決定係数が悪いのに、誤分類数は他の系列よりも良かった。このことは、多重共線性の問題点を視覚的に示す例と考える。

表2 AIC最小モデルにおける誤分類数と誤分類されたケース

	誤分類	誤分類されたケース
IP	10	38,58,113,174,204,206,207,218,220,227
LP	19	15,30,38,58,63,90,103,113,130,174,204,207,214,215,216,220,227,236,239
F(3:1)	17	15,30,38,47,58,63,90,103,113,130,174,204,206,207,210,220,227
Q(3:1)	18	15,30,38,58,63,90,103,113,130,174,204,206,207,214,218,220,227,232
F(1:1)	22	4,15,22,30,34,38,47,51,58,63,87,88,90,95,103,113,129,130,141,158,174,204
Q(1:1)	18	15,22,30,34,38,47,58,63,88,90,95,103,113,129,130,174,204,220

表1は、AIC最小化基準で選ばれた4変数モデルの判別係数である。線形判別とそれほど異なっているわけではない。表2は、各手法で誤分類されたケース番号である。IP-OLDFの誤分類されたケースは、ほぼ他の手法に含まれている。F(3:1)は、事前確率を3対1に設定した。

図1は、線形判別の誤分類数をIPの誤分類数で回帰したものである。回帰式は、

FP=12.843+0.469IP

であり、相関係数は 0.84 と高かった。

図2は、QPをIPで回帰した例である。回帰式は、QP=11.172+0.912IPであり、相関係数は0.543と低い。2次判別関数は、推定パラメーターも多く、データに過敏に反応するからでなかろうか。

LP-OLDFを回帰すると、LP=4.776+1.316IPで、相関係数も0.861と高かった。

図1 線形判別の誤分類数をIPで回帰する

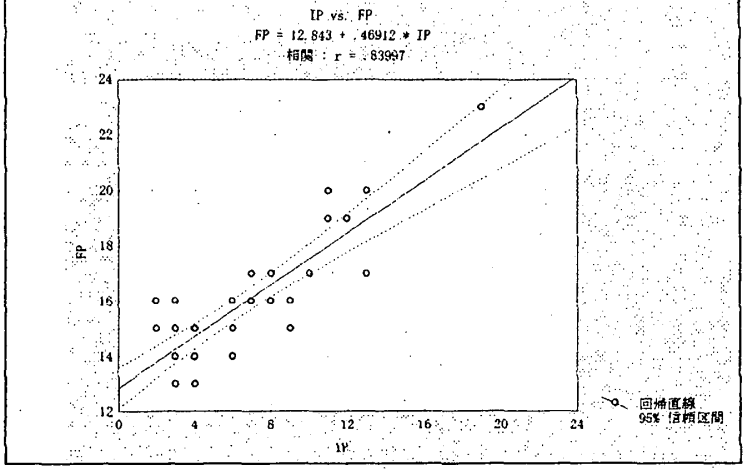
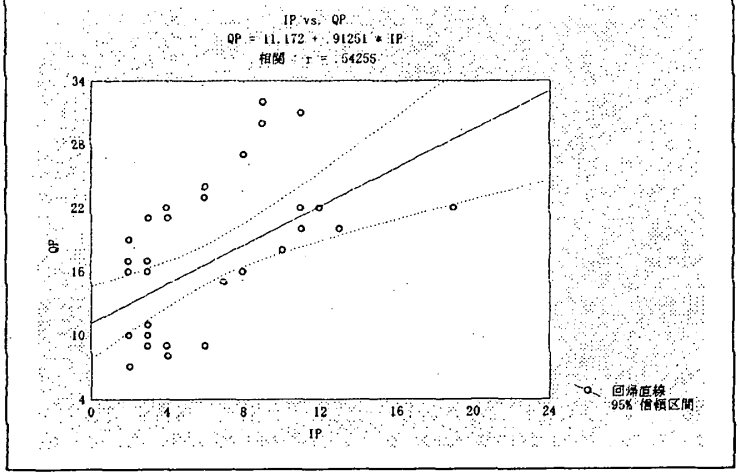


図2 QPをIPで回帰する



5 まとめ

IP-OLDF を現実の医学データに適用し、多重共線性の解消の効果や、従来の手法と比較評価し好成績を得た。

文献)

1)新村秀一(1998). 最適判別関数(2), 第66回日本統計学会大会, 165-166.

2)新村秀一(1998). 最適線形判別関数によるモデル決定, 日本行動計量学会第26回大会論文集、117-118.

3)三宅章彦・新村秀一(1980). 最適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、5-20.

4)A.Miyake,S.Shinmura(1976) .Error rate of linear discriminant function,North-Holland Pub.Company.435-445.

5)Glover, F(1990). Improve Linear programming models for discriminant analysis.Decision Sciences,2, 771-785.

6)新村(1996). 重回帰分析と判別分析のモデル決定(2)-19 変数をもつ CPD データのモデル決定-. 成蹊大学経済論集、第 27 巻第 1 号、180-203.