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

乱数データを用いた最適線形判別関数(OLDF)の評価

1202720 成蹊大学 新村秀一 SHINMURA Shuichi

1 はじめに

これまで、フィッシャーのあやめのデータと医学データを用いて、整数計画法を用いた最適線形判別関数(IP-OLDF)の評価を行ってきた。そこで得られた知見を、乱数データを用いて検証したい。

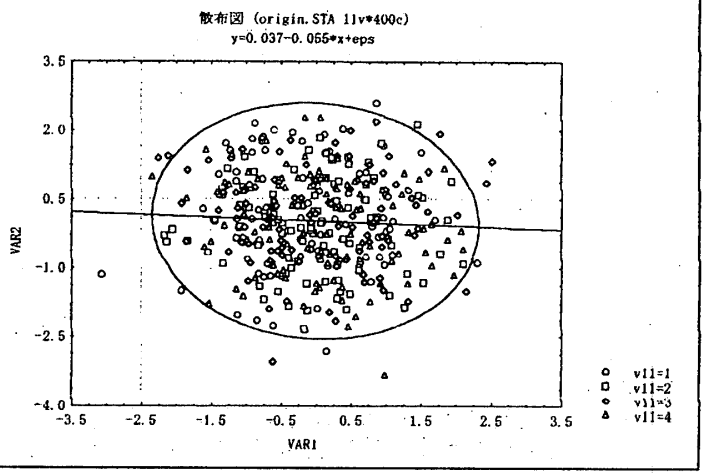
2 乱数データ

Speakeasy(文献2)を用いて、2変数の正規乱数x＝N(0,1)*2とy＝N(0,1)を400個作成した(図1)。それを、100件*2変数の4組の2群判別データとした。2組を内部標本G1とG2として、残り2組をそれらに対応する外部標本G3とG4とした。

さらに、G1群は、0度、30度、45度、60度、90度で回転した。G2群は、xに0から8までの整数を、yに0、2、4を加えて、平行移動した。このようにして作られる5*9*3組み合わせから、115組の2群判別の内部標本(G1とG2)と外部標本(G3とG4)を作成した。

例えば、D12A30はG2群のxとyに1と2を加え、G1群を30度回転した内部標本と外部標本のデータを表す。フィッシャーの線形判別関数は、DijA0のデータで、理論的前提を満たしていると考えられる。

図1 Speakeasy による乱数データ



3 解析結果

本データを、IP-OLDF、LP-OLDF、線形判別分析、2次判別分析で分析した。乱数データであるので、事前確率は、0.5 対 0.5 で問題はないだろう。

表 1 は、誤分類数の基礎統量である。FIT はフィッシャーの内部標本での判別結果である。FET はフィッシャーの外部標本、OIT は OLDF の内部標、OET は OLDF の外部標本、QIT は 2 次判別の内部標本、QET は 2 次判別の外部標本での判別分析の誤分類数を表す。内部標本の中央値を見ると、OIT<QIT<FIT になっている。最適線形判別が、2 次判別より良いことは注目に値する。しかし、外部標本では OET=QET <FET である。

表 1 4組のデータの基礎統量

	平均	Q2	最小	最大	Q1	Q3	標準偏差	標準誤差	歪度	標準誤差	尖度	標準誤差
FIT	17.017	9	0	71	4	27	17.620	1.643	1.250	0.226	0.622	0.447
FET	22.165	15	1	80	6	36	20.455	1.907	1.089	0.226	0.282	0.447
OIT	13.435	6	0	60	2	22	15.824	1.476	1.274	0.226	0.588	0.447
OET	21.939	14	1	84	7	33	19.748	1.841	1.168	0.226	0.502	0.447
QIT	15.678	8	0	64	3	26	16.594	1.547	1.122	0.226	0.120	0.447
QET	20.122	14	0	79	5	32	19.006	1.772	1.023	0.226	-0.019	0.447
FITOIT	3.583	3	0	17	2	5	2.800	0.261	1.668	0.226	4.755	0.447
QITOIT	2.243	2	-3	13	1	3	2.455	0.229	1.020	0.226	2.444	0.447
FETOET	0.226	0	-8	8	-2	2	3.121	0.291	0.200	0.226	0.081	0.447
QETOET	-1.817	-1	-15	9	-3	0	3.259	0.304	-0.842	0.226	3.072	0.447

FITOIT は、FIT から OIT を引いたものである。線形判別に比べ、IP-OLDF は 3.6 例ほど誤分類数が少ないことが分かる。平均値の標準誤差が 0.261 なので、t値は 13.7 となり棄却される。中央値は 3 で、範囲は 17 で、四分位範囲は3で、標準偏差は 2.8 である。歪み度は 1.668 で、標準誤差が 0.226 なので、右に

歪んだ分布である。尖がり度は 4.755 で、標準誤差が 0.447 なので、かなり右に裾を引いた分布であることがわかる。

QITOIT の平均値は 2.243(標準誤差 0.229)で、中央値は2である。2 次判別関数より、2 例誤分類数が少ないことが分かる。最大値は 13 例で、最小値は－3 例である。歪み度と尖り度から、右に裾を引いた分布であることが分かる。

FETOETの平均は0.226で、標準誤差が0.291なので、線形判別とIP-OLDFの外部標本の誤分類数に差がないと考えられる。中央値も0である。最大値と最小値は8と－8であり、Q3とQ1は2と－2である。歪み度と尖り度は、検定で0と考えられるので、正規分布と考えてもよさそうだ。

QETOETの平均は－1.817で標準誤差が0.304なので、負と考えられる。すなわち内部標本では、IP-OLDFが良かったが、外部標本では2次判別のほうが－1.8例ほど成績がよい。中央値－1である。歪み度と尖り度から、少し左に裾を引いた分布であることがわかる。

図2は、FITをOITで回帰したものである。回帰式は、

FIT=2.181+1.104OIT

であり、相関係数は 0.991 と高い。従来の判別関数の誤分類数は、事前確率やリスクの導入で誤分類数が異なってくる。IP-OLDF の利点の一つは、誤分類数がユニークに決まるので、各種判別手法の成果を比べる基準に用いることができる。

図2 FIT を OIT で回帰する

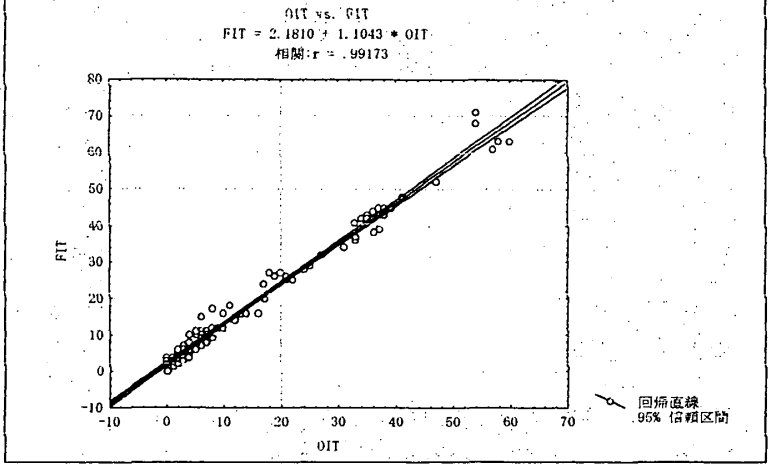


図3は、QITをOITで回帰している。回帰式は、

QIT=1.735+1.038OIT

である。相関係数は0.990である。CPDのような現実のデータと異なり、乱数データであるので、それほど当てはまりも悪くない。

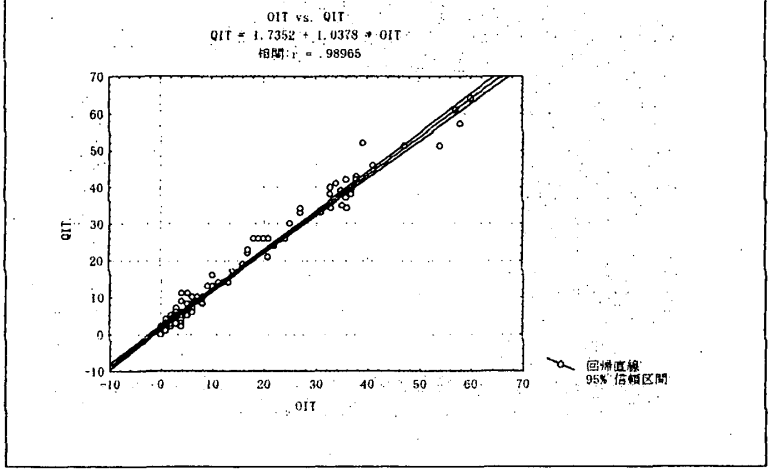
他の回帰分析は、次の通りである。

FET=5.122+1.269OIT

OET=5.430+1.229OIT

QET=4.399+1.170OIT

図3 QIT を OIT で回帰する



4 結論

IP-OLDF の誤分類数は、説明変数が増加するにつれ単調減少し、データに対しユニークにきまる。今回は、CPD データで得られた結果を乱数データで確認することにした。

今後の課題として、判別分析における多重共線性の問題点を乱数データで確認したい。

文献)

1)新村秀一(1999)。整数計画法を用いた最適線形判別関数(OLDF)、1999年度オペレーションズリサーチ学会秋季大会。

2)新村秀一(1999)。パソコンらくらく数学、講談社。