連載講座

ORと統計の融合

数理計画法を用いた最適線形判別関数(4)
—2変量正規乱数データによるIP-OLDFの評価—

新村　秀一

今回は，115組の2変量正規乱数データを用いて，アイリスデータと医学データの結果との比較及びExternal Checkを行った．External Checkやクロスバリデーションは，分析結果を現実問題へフィードバックするために必要な検討事項である．

内部標本とは，判別式を作成するために用いたデータのことである．その判別結果を検討することをInternal Checkという．これに対して，計算に用いていないデータ（外部標本）に判別式を適用し，その判別結果を検討することをExternal Checkという．最近では，クロスバリデーションということが多い．

1．2変量正規乱数データ

(1) 正規乱数データの発生

正規乱数データを，オブジェクト指向言語Speakeasy（新村（1999））を用いて，以下のように作成した．

```
X=NORMRANDOM (ARRAY (400, 1:))*2;
Y=NORMRANDOM (ARRAY (400, 1:));
```

400行×1列の2組の配列に，平均0で標準偏差が2と1の正規乱数を発生して，それを配列XとYとした．このデータを，400件×2変数からなる2変量正規乱数データとする．そして，これを100件×2変数からなる，4組の2群判別用の元データとした．最初の2組を内部標本G1とG2として用い，残り2組をそれらに対応する外部標本G3とG4とする．すなわち，G1群は1番目から100番目のケースである．G3群は201番目から300番目のケースであり，G1群の外部標本と考えている．同様に，G2群は101番目から200番目のケースである．G4群は301番目から400番目のケースであり，G2群の外部標本と考える．

(2) 判別データの作成

さらに，この4組の元データから，115組のデータを次のように作成した．

G1群とG3群のデータは原点を中心に，0度，30度，45度，60度，90度のそれぞれで回転を行った．G2群とG4群のXの値には0から8までのいずれかの整数iを，Yには0，2，4のいずれかの整数jを加えて，原点を（i, j）に並行移動した．

Speakeasyを用いたのは，データの回転や移動が配列や行列演算として簡単に行えるからである．

このようにして作られる5×9×3個の組み合わせ（135組）から，誤分類率が0になるものや50%近くになるもの20組を省いた残り115組を本研究のためのデータとした．すなわち，115組の内部標本（G1とG2）と，それに対応する外部標本（G3とG4）を作成したことになる．

この1組の判別データを，DijAkと表すことにする．DijAkは，G1群とG3群をk度回転し，G2群とG4群の原点を（i, j）に平行移動したものを表す．例えば，D12A30はG2群とG4群のXとYに1と2を加え，G1群とG3群を30度回転した内部標本と外部標本の4群のデータを表す．

G1群とG2群で各判別関数を適用し（Internal Check），その判別式を用いてG3群とG4群を判別（External Check）するものとする．

Di0A0タイプのデータは，2群が正規分布で等分散であるので，線形判別関数は良い結果が得られると考えられる．そして，Di0Akのkをきくすると，2次判別関数やIP-OLDFは，線形判別関数に比べて良くなることが期待される．また，jを2と4にして平行移動しても，等分散性が棄却されることになる．Shinmura & Miyake（1979）では，ヒューリスティックなOLDFを用いて，Di0A0タイプのデータを分析し，内部標本ではヒューリスティックなOLDFが線形判別関数よりよく，外部標本ではヒューリスティ

しんむら　しゅういち
成蹊大学 経済学部
〒180-8633 武蔵野市吉祥寺北町3-3-1

ックなOLDFの方が悪くなるという結果が得られている．

回転と平行移動を組み合わせることで，多くのテストデータが簡単に作成でき，2次元における2群のデータに関して布置のかなりのパターンをカバーできたと考える．

2. 誤分類数による評価

(1) 分析結果

本データを，4つの判別関数（Fisherの線形判別関数，2次判別関数，LP-OLDF，IP-OLDF）で分析し，その誤分類数で評価する．乱数データでありデータ件数が等しいので，事前確率は0.5対0.5として問題はないだろう．また，リスクを考慮する必要もないだろう．

表1に115組の判別結果のうち，10組を掲載した．最初の列は判別データ名を，XとYはG2群とG4群のXとYに加えた値（i, j）を，DEGはG1群とG3群を回転させた角度kを表す．FIT以降の8変数は，各判別分析の誤分類数である．

FITは，線形判別関数の内部標本での誤分類数である．FはFisherの線形判別関数を，ITは内部標本（Internal Sample）を表す．FETのETは外部標本（External Sample）を表す．LITはLP-OLDFの内部標本を，LETは外部標本の誤分類数を表す．OITはIP-OLDFの内部標本，OETは外部標本での誤分類数を表す．QITは2次判別関数の内部標本，QETは外部標本の誤分類数を表す．

その後に続く3個の変数は，これらの誤分類数を比較するために，各判別手法の誤分類数の差を求めた．例えば，FITOITは，FITからOITを引いたものを表す．これは，IP-OLDFが線形判別関数に比べてどれだけ誤分類数が少ないかを表している．

(2) 回転の影響

G1群を回転させることにより，Di0Akタイプのデータでは，線形判別関数よりも2次判別関数やIP-OLDFの方が良くなると考えられる．すなわち，kが大きくなるにつれFITOITとFITQITは値が大きくなると考えられる．

太字で示したD10Akの5個の結果は，回転するにつれ，FITOITの誤分類数の差は3例から17例へ単調に増えている．D20Akでは，D20A90を除けば，5から8へ単調に増えている．このような明らかな傾向は，誤分類数も少なくなることも影響して，他のタイプのデータでは確認できない．

FITQITは，D10Ak，D20Ak，D70AkとD80Akで，回転するにつれ2次判別関数の方が線形判別関数より成績が単調に良くなっている．この点で，2次判別関数の方がIP-OLDFより傾向が明らかである．

以上から，一般に期待される回転の影響は，IP-OLDFより2次判別関数で認められたが，全てのDi0Akタイプのデータで認められなかった．すなわち，本データでは，回転により等分散でなくなっても，あまり大きな影響を及ぼさないようだ．

(3) 平行移動の影響

同じXの値であっても，Yが0，2，4に対応して誤分類数が急に少なくなることがレーダーチャートを調べることで分かる．そして，おなじYの値の場合，Xが大きくなるにつれ，誤分類数は減少している．

3. 主成分分析とクラスター分析

次に，得られた誤分類数を総合的に把握するため，主成分分析とクラスター分析で検討する．

(1) 主成分分析

FITからQETまでの8個の変数で主成分分析を行った．第1主成分の寄与率が98.4%であり，ほぼ8個の情報は第1主成分で表される．図1で，第1と第

表1 誤分類数

Group	X	Y	DEG	FIT	FET	LIT	LET	OIT	OET	QIT	QET	FITOIT	QITOIT	FITQIT
D10A0	1	0	0	63	80	80	83	60	84	64	79	3	4	-1
D10A30	1	0	30	61	78	69	81	57	79	61	64	4	4	0
D10A45	1	0	45	63	78	63	81	58	72	57	62	5	-1	6
D10A60	1	0	60	68	72	66	81	54	66	51	56	14	-3	17
D10A90	1	0	90	71	69	72	81	54	64	51	54	17	-3	20
D20A0	2	0	0	52	66	66	71	47	68	51	66	5	4	1
D20A30	2	0	30	48	60	61	71	41	60	46	54	7	5	2
D20A45	2	0	45	45	56	55	71	37	52	40	45	8	3	5
D20A60	2	0	60	43	50	55	72	35	42	35	39	8	0	8
D20A90	2	0	90	42	41	46	70	36	43	34	39	6	-2	8

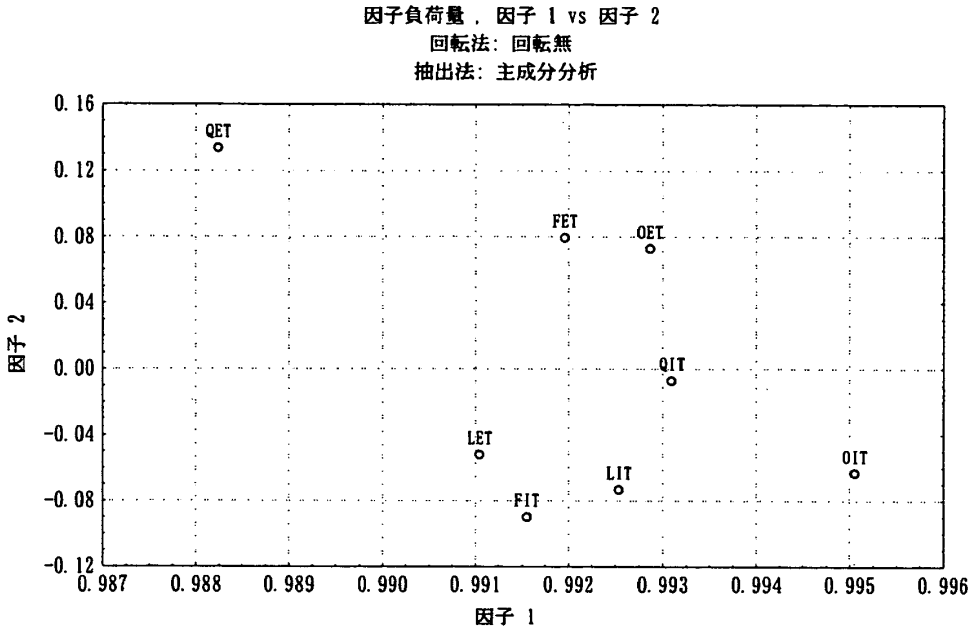

図1　因子負荷量

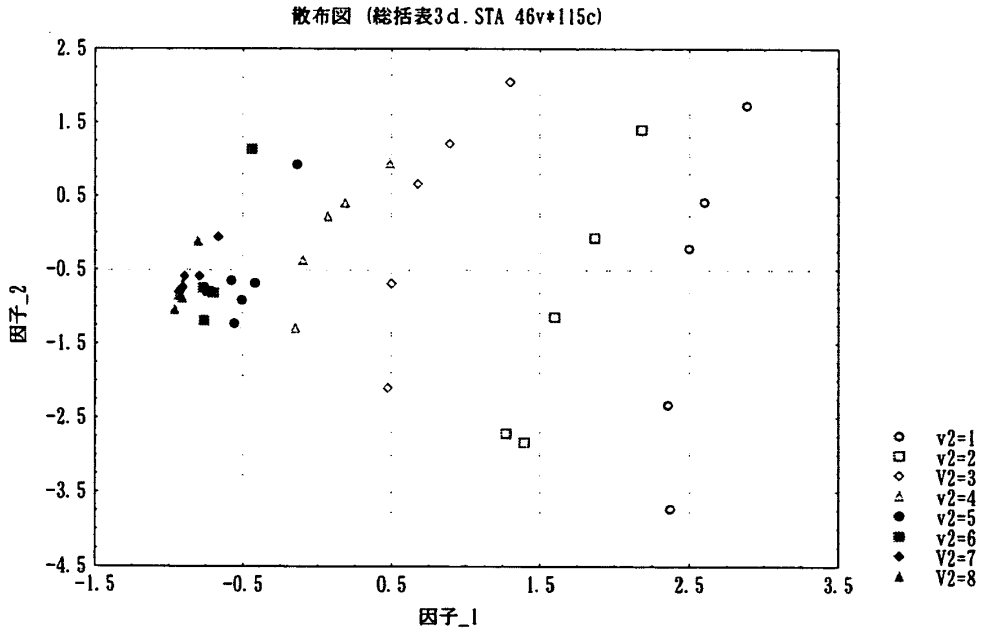

図2　サンプルスコア

2因子の負荷量をプロットしてみると，外部標本に関するQET，FET，OETが第1象限にあり，LETと内部標本の4変数は第4象限にあることが分かる．第2主成分は，ほとんど情報を持っていないが，一応内部標本と外部標本を分ける軸のようだ．第1軸では，OITとQETと残りの6変数の3グループに分かれていることが分かる．すなわち，内部標本のIP-OLDFの誤分類数が少なく，外部標本の2次判別関数の誤分類数が他に比べて多いことを表している．第1因子軸は誤分類数の少なさを表す総合特性値と考えてよさそうだ．

図2は，Y＝0のサンプルスコアである．第1因子軸の正(3)から負（−1）にかけて，D10Ak（V2＝1, ○）からD80Ak（V2＝8, ▲）のタイプに対応している．(i, 0) が同じグループでkが大きくなると，第2因子軸の値は正から負になる．iが大きくなると，回転による影響（第2因子軸のバラツキ）は小さくなり，第1因子軸での各グループ間の間隔も小さくなっている．すなわち，D50AkからD80Akは重なり合って区分が難しい．

Y＝2のサンプルスコアからD02Ak，D12Ak，D22Ak，D32Akは，一応区別ができるけれど，第1因子負荷量の値の違いで整然と区別できない．D42AkからD72Akは，研究論文では乱数データによる評価に重きをおく傾向があるが，異なったグループとして区別できないことが分かった．

Y＝4のサンプルスコアから第1因子軸の範囲が[−1.1，−0.1]と狭く，D04AkからD64Akの7個のグループはさらに重なり，区別できないことが分かった．

以上から，Yを2と4へ平行移動すると，回転の影響が小さくなることが分かる．

(2) クラスター分析

最近隣法，最遠隣法，重み付き平均法（図3）などの手法によって異なった変数のクラスターが得られた．その主な原因は，LP-OLDFが異なるクラスターに入ることによって起こっている．

この点を除けば，内部標本と外部標本同士で大きなクラスターを構成している．その中で，内部標本では2次判別関数とIP-OLDFが，外部標本では線形判別関数とIP-OLDFがサブクラスターを形成しているという特徴を持っている．

以上から，LP-OLDFは，計算時間が少なくIP-OLDFの代替手法になるかと期待されたが，現実の適用には問題があるようだ．

4. 基本統計量の検討

(1) 誤分類数の基本統計量

表2は，誤分類数の基本統計量である．ケース数は115例である．

例えば，FITは0から71の値を取り，内部標本の誤分類率は0%から35.5%になる．誤分類率が大きなデータを判別しても応用上あまり意味がないので，本データは判別手法の分析データとしては適切と考える．平均値は17.017であり，中央値は9であり，歪み度が1.250（標準誤差は表より省略してあるが0.226）なので，右に裾を引いた分布である．尖り度から，それほど大きな外れ値はない．四分位範囲は23であり，標準偏差は17.620である．

(2) 中央値の比較

中央値の大小を比較すると，OIT(6)＜QIT(8)＜FIT＝LIT(9)＜LET(13)＜OET＝QET(14)＜FET(15)になる．

内部標本の全ての中央値は9以下であり，外部標本は13以上である．内部標本の誤分類数が，外部標本の中央値より悪ければ問題であるが，それはなかった．内部標本では，IP-OLDFの方が，2次判別関数より成績が良いことは注目に値する．2次判別関数は，推定パラメータが多いので，IP-OLDFや線形判別関数より成績が一般的に良くなると思われる．その反面，外れ値などのデータの影響を受けやすいと考えられる．

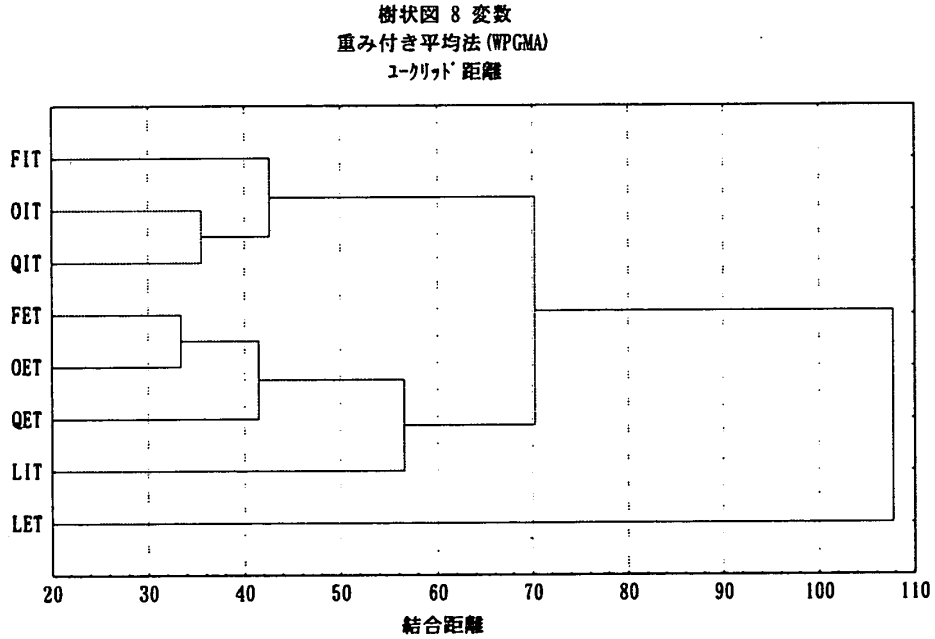

図3 変数のクラスター分析

表2　基礎統計量

	平均	中央値	最小値	最大値	25%点	75%点	標準偏差	標準誤差	歪み度	尖り度
FIT	17.017	9	0	71	4	27	17.620	1.643	1.250	0.622
FET	22.165	15	1	80	6	36	20.455	1.907	1.089	0.282
LIT	18.713	9	0	80	3	32	20.757	1.936	1.163	0.162
LET	24.878	13	1	83	7	42	24.279	2.264	1.039	-0.312
OIT	13.435	6	0	60	2	22	15.824	1.476	1.274	0.588
OET	21.939	14	1	84	7	33	19.748	1.841	1.168	0.502
QIT	15.678	8	0	64	3	26	16.594	1.547	1.122	0.120
QET	20.122	14	0	79	5	32	19.006	1.772	1.023	-0.019
FITOIT	3.583	3	0	17	2	5	2.800	0.261	1.668	4.755
QITOIT	2.243	2	-3	13	1	3	2.455	0.229	1.020	2.444
LITOIT	5.278	3	0	20	1	8	5.485	0.511	1.185	0.426
FITQIT	1.339	1	-7	20	0	2	3.156	0.294	3.083	15.286
LITFIT	1.696	0	-6	17	-1	3	4.575	0.427	1.262	1.060
LITQIT	3.035	1	-4	21	-1	5	5.346	0.498	1.474	1.457

この点で，IP-OLDFは，推定パラメータも少なく，誤分類数も少ないので，2次判別関数より好ましいことは明らかである．

2次判別関数は，線形判別関数やLP-OLDFより良かった．この点は，医学データと異なっている．乱数データであるためであろう．一方，線形判別関数とLP-OLDFの誤分類数の中央値は，同じである．この点からも，あえてLP-OLDFを新手法として薦めることはできない．

外部標本では，LP-OLDFの成績が一番良く，IP-OLDFと2次判別関数の中央値が同じ値になった．線形判別関数のそれが一番悪かった．しかし，LETの第3四分位数（75%点）と平均値は，FET，OET，QETに比べて一番大きかった．これは医学データでも指摘されたが，LP-OLDFの誤分類数の振幅の幅が大きいことを示しているものと考えられる．

(3)　平均値の比較

平均値の大小順は，OIT(13.4)＜QIT(15.7)＜FIT(17)＜LIT(18.7)＜QET(20.1)＜OET(21.9)＜FET(22.2)＜LET(24.9)になる．LET以外では，中央値の結果とそれほど大きな違いはない．

表3は，平均値の差の検定結果を表すt値である．OETとFET以外は全て1%で棄却された．

以上から，次のように分かりやすい結論が得られる．内部標本でも外部標本でも，IP-OLDF＜線形判別関数＜LP-OLDFの順に成績が悪くなる．内部標本では，2次判別関数はIP-OLDFより悪く，外部標本では2次判別関数がIP-OLDFよりも良かった．

すなわち，アイリスデータや医学データと異なり乱数データでは，2次判別関数が健闘している．このことは乱数データを用いている限り，推定パラメータが

表3　t検定

	FIT	FET	LIT	LET	OIT	OET	QIT	QET
FIT	0.00	-11.68	-3.97	-11.05	13.72	-11.58	4.55	-6.82
FET	11.68	0.00	8.24	-4.91	16.17	0.78	12.99	6.29
LIT	3.97	-8.24	0.00	-13.23	10.32	-8.39	6.09	-3.06
LET	11.05	4.91	13.23	0.00	13.41	4.91	11.28	7.07
OIT	-13.72	-16.17	-10.32	-13.41	0.00	-18.24	-9.80	-14.19
OET	11.58	-0.78	8.39	-4.91	18.24	0.00	14.93	5.98
QIT	-4.55	-12.99	-6.09	-11.28	9.80	-14.93	0.00	-11.31
QET	6.82	-6.29	3.06	-7.07	14.19	-5.98	11.31	0.00

$p(p-1)/2$個も多いことによる問題点が明らかにならないことを示唆していると考えた方が良さそうだ．線形判別関数は，内部標本でも外部標本でも，IP-OLDFと2次判別関数より悪かった．

(4)　XとYによる層別箱ひげ図による検討

図4は，OITのXとYによる層別箱ひげ図である．3つのグラフの左上，右上，左下は，Y＝0，2，4に対応している．各グラフの横軸はXの0から8に対応し，縦軸は誤分類数を表している．個々の箱ひげ図は，G1群の角度を5個回転させたものが描かれている．Y＝0の場合（左上のグラフ），Xの値が増えると急速に誤分類数は減少している．これに対して，Y＝2の場合（右上のグラフ），X＝0と1でそれほど減少しないが，2以上で減少している．Y＝4の場合は，緩く減少している．これを見れば，G1群の回転よりも，G3群の平行移動の方が誤分類数に大きな影響を与えていることが分かる．また，誤分類数が少なくなるにつれ，回転の影響も少なくなっている．

他の7個の変数も程度の違いはあるが，似たような傾向を示した．

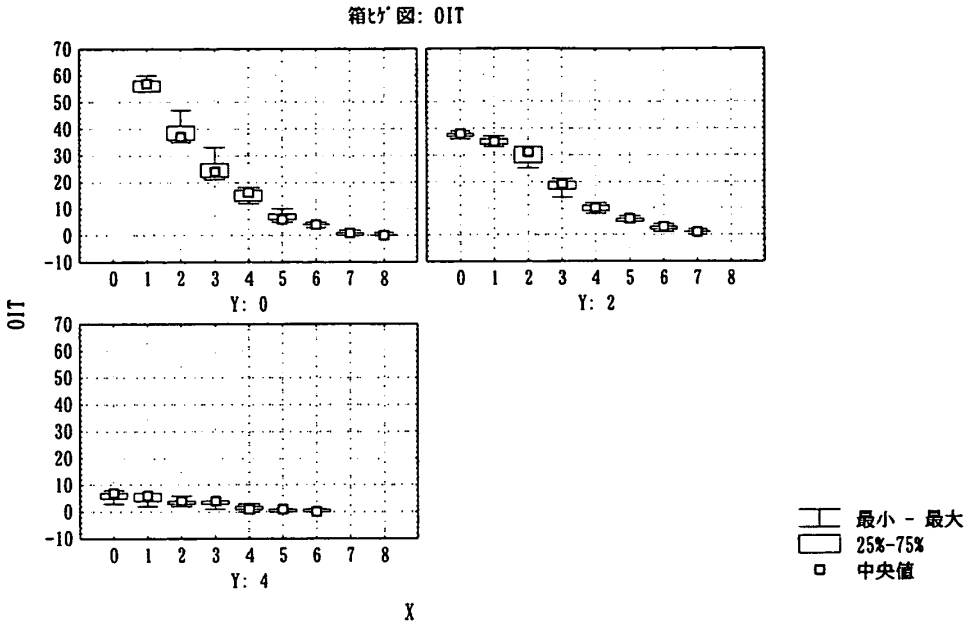

図4　OITの誤分類数の箱ひげ図

5.　相関分析と回帰分析

(1)　相関係数の検討

相関分析を行ったところ，IP-OLDFと線形判別関数の相関係数が0.992と一番大きく，他の相関係数も0.967以上であった．行列散布図から，全て直線相関であることを確認した．そこで，IPの誤分類数で他の誤分類数を回帰分析し，評価することにする．

(2)　回帰分析による検討

誤分類数は，事前確率やリスクの値によって異なり，判別結果の比較評価に用いにくい．これに対して，IP-OLDFの誤分類数は一意に決まるので，それを説明変数として他の誤分類数を評価することが考えられる．

表4は回帰式と相関係数をまとめたものである．

FITをOITの誤分類数で回帰した回帰式は，FIT＝2.181＋1.104×OITであり，相関係数は0.992と高い．OITに比べ，10%ほど誤分類数が増えるようだ．

図5は，内部標本の回帰直線を表している．一番上の一点鎖線はLP-OLDF，実線は線形判別関数，破線は2次判別関数，一番下の太い破線はIP-OLDFをIP-OLDFで回帰した回帰直線を表す．これら4手法の誤分類数の予測値は，平均の大小順と同じく，IP-OLDF＜2次判別関数＜線形判別関数＜LP-OLDFになる．ただし，OITの値の小さなところではこれらの差もなく，大きくなるに従い差も大きくなることが分かる．

表4　回帰式

回帰式	相関係数
FIT=2.181+1.104×OIT	0.992
QIT=1.735+1.038×OIT	0.990
LIT=1.243+1.300×OIT	0.991
FET=5.122+1.269×OIT	0.981
QET=4.399+1.170×OIT	0.974
LET=4.594+1.510×OIT	0.984
OET=5.430+1.229×OIT	0.985

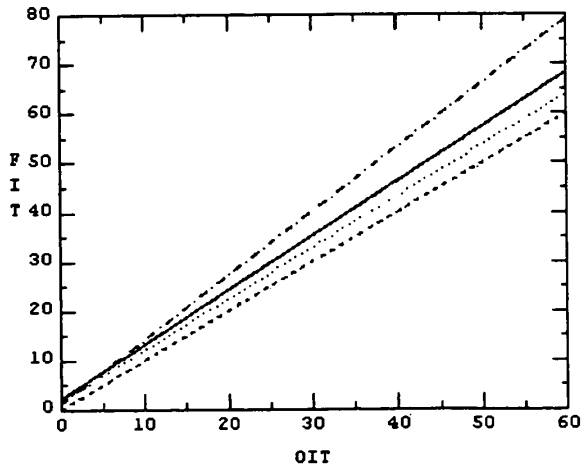

図5　内部標本の回帰式

図6は，外部標本の回帰直線を表している．OETをOITで回帰した直線は，実線で表した線形判別関数と破線の2次判別関数の間にくる太い破線である．この結果も，次のように平均値で検討した大小順と同

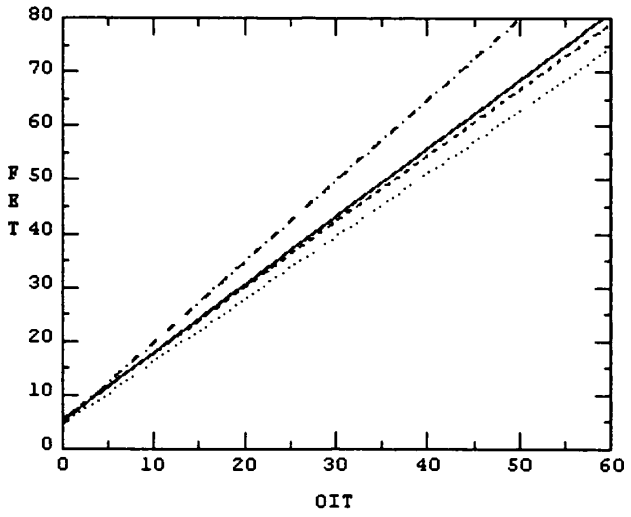

図6　外部標本の回帰式

じである.

2次判別関数<IP-OLDF<線形判別関数<LP-OLDF. そして, OITが大きくなるに従い, 手法間の違いが大きくなっている.

6. 結論

IP-OLDFの誤分類数は, データに対しユニークに決まり, 説明変数が増加するにつれ単調減少するという特徴がある. 今回は, アイリスデータと医学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.

その結果, 内部標本の平均値の大小順は, IP-OLDF<2次判別関数<線形判別関数<LP-OLDFの順であった.

現実のデータでは, 外部標本によるExternal Checkは行いにくい. 乱数データであるため, 今回これが行えた. その結果, 2次判別関数<IP-OLDF<線形判別関数<LP-OLDFの順であった.

一方, アイリスデータと医学データでは, IP-OLDF<LP-OLDF<線形判別関数<2次判別関数, という結果が得られている.

これらを突き合わせると, LP-OLDFは現実のデータでは既存の手法より良くなる場合があるが, 乱数データでは一番成績が悪かった. これに対して, 2次判別関数は現実のデータでは一番成績が悪かったが, 乱数データでは線形判別関数やLP-OLDFより良かった. これは, 乱数データであるためと考えられる.

すなわち, 研究論文では乱数データによる評価に重きをおく傾向があるが, 現実のデータと乱数データの両方で評価することが重要であることが分かる.

いずれにしても, 2次判別関数とLP-OLDFは, 乱数データと現実のデータの両方で評価することで, 適用に問題があることが指摘できた.

これに対して, IP-OLDFは, 乱数データと現実のデータの両方で好成績を得たことは, 評価できよう. 外部標本では2次判別関数より悪かったが, 線形判別関数より悪くなかったのは, 意外な結果である.

ヒューリスティックなOLDFでは, External Checkで線形判別関数より悪かったのは, この手法が大局的な最適解を求めていないか, データがいわゆるDi0A0タイプでしか評価しなかったことのいずれかであるが, 後者の可能性が大きい.

今後の課題として, さらに3変数以上の正規乱数データや, 他の分布による乱数データ, あるいは物理乱数データによる検討を行っていく必要がある. さらに, 医学データで得られた多重共線性に関する知見を検証する必要がある.

また, 多くの判別分析やクラスター分析の研究でアイリスデータが良く利用されているが, 優れて現実的なデータを発掘あるいは乱数データを設計し, 広く研究者に提供していくことが今後重要と考える.

参考文献

[1] S. Shinmura & A. Miyake, Optimal linear discriminant functions and their applications, Proceedings of the COMPSTAT 79, pp. 167-172, 1979.

[2] 新村秀一：パソコンらくらく数学, 講談社, 1999.

[3] 新村秀一：数理計画法を用いた最適線形判別関数, 計算機統計学, 11-2, pp. 93-105, 1998

[4] 新村秀一・垂水共之：2変量正規乱数データによるIP-OLDFの評価, 計算機統計学, 12-2, pp. 107-124, 1999.

[5] S. Shinmura, Optimal Linear Discriminant Function (OLDF) using Mathematical Programming, Bulletin of the International Statistical Institute ISI 99 52nd Session Contributed Papers Book 3, pp. 247-248, 1999.