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(2)
—アイリスデータへの適用—

新村　秀一

1. はじめに

これまで，多くの判別手法が提案されてきたが，現実問題への適用によってその有用性を示せず消えていったものも数多くある．その理由としては，既存の手法に比べて，操作性が悪い，計算時間がかかる，理論の前提が現実ばなれしている，判別成績が悪い，内部標本で成績が良くても外部標本で成績が悪い，等の理由が挙げられる．

判別分析では，Fisherの線形判別関数が既存の代表的な手法である．操作性に優れ，計算時間が速いことが特徴である．2次判別関数は，操作性や計算時間ではFisherの線形判別関数に比べて劣るが，2群が等分散でない場合，少なくともFisherの結果に比べて成績がよいことが理論上期待できる．これらの手法は，これまでの多くの統計手法と同じく，正規分布を前提に構築されている．しかし，多くの現実のデータは正規分布であることは少ないし，多重共線性などの問題があるため，これらの手法は全ての判別問題に適しているとは考えられない．

IP-OLDFは計算時間がかかるが，特定の分布を仮定していない点に特徴がある．すなわち，広範な現実のデータに適用可能である．また，Fisherの線形判別関数と共に用いることで，判別結果に違いがなければ対象データはFisherの線形判別関数の仮定に従い2群は多次元正規分布で等分散と考えてもよく，IP-OLDFの結果がFisherの線形判別関数よりも良ければこの仮定に従わないと考えられる．

今回は，これらの点を有名なアイリスデータで評価することにしたい．このデータは，新村（1997）のCD-ROMに収録している．また，http://sun.econ.seikei.ac.jp/~shinmura/からダウンロードできる．

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

2. アイリスデータ

アイリスデータは，3種のアイリス（セトナ・バーシクル・バージニカ）から，がく片（X1）・がく片幅（X2）・花びら（X3）・花びら幅（X4）という4変数を各50個ずつ計測したものである．

セトナは，他の2群と容易に判別できるので，本研究ではバーシクルとバージニカの2群を用いる．すなわち，各50個の2群を4個の説明変数で判別する問題である．この場合，事前確率はデータ数と同じ0.5対0.5と考えればよく，リスクも考える必要はない．

本データは，統計学の泰斗であり自ら線形判別関数にも名を残す英国のFisher卿が判別分析のテストデータに用いたことで，統計研究者の間でつとに有名である．このため，判別関数やクラスター分析の研究や説明に必ずといってよいほど用いられている．

この点で，多くの統計研究家が知識を共有するこのデータで，IP-OLDFを評価することに意味がある．またIP-OLDFは，線形で表される判別関数の中で標本誤分類率が一番良いので，本講座の結果は今後開発される判別関数の手法の評価に目標を与えることになる．

一方，これほど有名なデータであるにもかかわらず，基本統計量から始まって徹底的に解析されてこなかった．新村（1997）は，データ解析の教科書として，このデータを徹底的に分析している．またこのデータは，相関係数の間違った解釈を避けるという統計教育においても重要な意味を持っている．

3. 解析結果の説明

表1は，SASによる全ての回帰モデル（総当り法）の出力結果に，今回提案するIP-OLDFとLP線形判別関数，Fisherの線形判別関数と2次判別関数の判別結果を付け加えたものである．

最初の数字(pの列)は，判別分析(回帰分析)に用いられる説明変数の数を表わしている．4個の説明変数

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

表1　総当り法による全ての判別（回帰）モデル[1]

p	R-square	C(p)	AIC	IP	LP	F5	Q5	Independent Variables
1	0.686	42.1	-250	5	5	6	6*	X4
1	0.618	71.7	-231	5	5	8	7	X3
1	0.244	236.2	-163	24	24	27	30	X1
1	0.095	301.9	-145	29	35	42	42	X2
2	0.724	27.4	-261	3	7	5	7*	X2 X4
2	0.720	29.2	-260	3	4	6	3**	X3 X4
2	0.697	39.1	-252	3	4	6	6	X1 X3
2	0.686	44.1	-248	5	6	6	5**	X1 X4
2	0.632	67.6	-233	5	5	7	10	X2 X3
2	0.246	237.4	-161	24	24	25	29	X1 X2
3	0.767	10.4	-276	2	2	4	4**	X2 X3 X4
3	0.760	13.6	-273	2	2	3	3**	X1 X3 X4
3	0.729	27.2	-261	3	6	5	6*	X1 X2 X4
3	0.700	40.1	-251	2	3	7	8	X1 X2 X3
4	0.784	5.0	-282	1	1	3	3**	X1 X2 X3 X4

から，15個の判別分析のモデルが作られる．最右端の変数はモデルに含まれる説明変数を表わす．2番目以降の数字は，決定係数，Mallow'sのCp統計量，赤池のAIC基準である．ここまでがSASの出力である．

それ以降は，IP-OLDF（IP），LP線形判別関数（LP），線形判別関数（F5）と2次判別関数（Q5）による誤分類数を表わしている[1]．F5とQ5の5は，事前確率が0.5対0.5であることを表わしている．

例えば，1行目は，X4を説明変数とするモデルである．バーシクルに0とバージニカに1というダミーの目的変数値を与えた回帰分析の結果として，そのモデルの決定係数が0.686，Mallow'sのCp統計量が42.1，赤池のAIC基準が−250になった．各判別関数による誤分類数は，それぞれ5個，5個，6個，6個である．このモデルは，説明変数が1個のモデルの中で，決定係数が1番よいモデルである．モデル（X2）は，決定係数が0.095で一番成績が悪いモデルである．

pが2すなわち説明変数が2個のモデルは，全部で${}_4C_2=6$個あり，その中でモデル（X2，X4）の決定係数が一番良いことを表わす．

2次判別関数の誤分類数につけられた＊や＊＊は，2群の分散共分散が等しいか否かのχ^2検定が，5%か1%で棄却されたことを示す．この場合，理論上は2次判別関数を用いることが望ましいことになる．

しかし，χ^2検定で棄却された8モデルでは，5%で棄却されたモデル（X2，X4）と（X1，X2，X4）の2例がFisherの線形判別関数の方が2次判別関数より良く，5%と1%で棄却された4モデルが同じであり，1%で棄却された2モデルだけで2次判別関数の方が線形判別関数より良かった．結局χ^2検定は，4月号で紹介する管理された乱数データでしか良い結果がでないのではないかと考えられる．

4. 逐次変数選択法と総当り法によるモデル決定

4.1 総当り法と基本系列

表1のように，総当り法の出力があれば，逐次変数選択法などをこの表の上でシミュレーションできる．

逐次変数選択法の変数増加法で停止則（F_{in}基準）を考えないで，1変数からフルモデルの4変数まで求めたものを上昇基本系列と呼ぶことにする．表1から，このモデル系列は，（X4）→（X2，X4）→（X2，X3，X4）→（X1，X2，X3，X4）であることが分かる．変数減少法で停止則（F_{out}基準）を考えないで求められた下降基本系列は，（X1，X2，X3，X4）→（X2，X3，X4）→（X2，X4）→（X4）である．すなわち，このデータでは上昇基本系列と下降基本系列で選ばれたモデルは一致している．

両方のモデル系列が完全に一致し，それが各p次元で一番良いモデルである場合，最終的に選ぶモデルはこのモデル系列から選ぶことは理にかなっている．しかし，不一致の場合は，モデル決定は難しくなる．

多くの統計ソフトでは，逐次変数選択法を提供しているが，総当り法が利用できない場合が多い．そのような場合は，逐次変数選択法でこの両系列のモデルを慎重に検討すればよい．

4.2 モデル決定

良いモデルを決定することに関して，オーソライズされた意見はまだない．しかし，良いモデルは，基本系列に含まれることが重要である．もし，上昇基本系列と下降基本系列が一致しておれば，これに含まれるモデルはいずれも各pで決定係数が一番良いモデルになっていることが多いので，このモデル系列からモデルを選ぶ理由は大きい．

本研究では，これまでの経験から，次のように考えている．

①1変数からフルモデルにおいて，上昇と下降基本系

[1] p変数モデルで，判別境界に（p+1）以上のデータがある場合，IP解の修正が必要である．アイリスデータではp=2でこれが起きているので，散布図で確認し行った．pが3以上の場合の一般的な方法は見つかっていない．

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

列で，Mallow'sのCp統計量や赤池のAIC基準を参考にして，適切な説明変数の個数pを決める．本論文では，AIC最小基準を用いる．このためには，停止則を考えることは意味がない．

②このpあるいはそれより少ない説明変数のモデルの中から，決定係数やMallow'sのCp統計量や赤池のAIC基準，あるいはその他の判別結果を参考にして，統計的に妥当なモデルを決める．多重共線性などで，基本系列が異なる場合，多重共線性を解消して再検討するか，基本系列にないモデルも検討対象にする必要があろう．

③固有領域の研究者にこの選んだモデルを提示し，彼らの考えと調整し，必要であれば再検討する．すなわち，統計的に良いモデルをできるだけ早く選んで，それを現実に役立つモデル選択の出発点にすべきである．

以上のおおまかなモデル選択の手順に，大きな問題はないであろう．

このような問題点を提起し知識を共有できる点が，アイリスデータの存在意義である．その反面，4変数しかないこと，誤分類確率がたかだか数%しかないので，各判別手法の違いが僅かであり，研究データとしては物足りないことは明らかである．

5. 誤分類数の比較

基本系列上のモデルの誤分類数を図1で示す．IP-OLDFの誤分類数は，5，3，2，1と単調に減少する．これに対し，LP線形判別関数では，2変数で7に増えて，後はIP-OLDFと一致している．F5は6，5，4，3と単調に減少しているが，Q5では2変数で7と増え，後はF5と一致している．

すなわち，LP線形判別関数と2次判別関数は，特定のデータに影響され2変数で判別成績は悪くなっている．このように，わずか4変数しかないが判別結果に違いがでる点でも興味ある問題を提供している．しかし，誤分類数の差がわずかであるので，それでもって明確な優劣の比較を行えないという欠点もある．

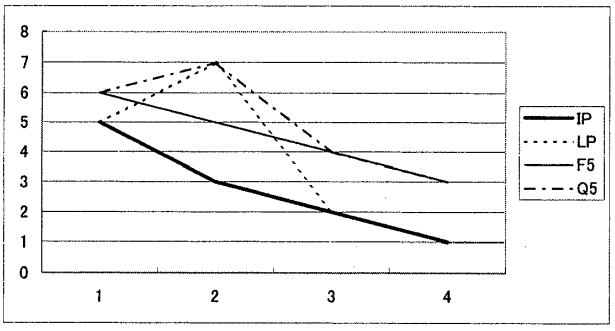

図1　基本系列モデルの誤分類数

6. 回帰分析による検討

6.1 回帰式による評価

表1の15例のLP，F5，Q5をIPの最小誤分類数で回帰すると，次の回帰式が得られた．

$LP=0.589+1.088\,IP$

$F5=1.593+1.173\,IP$

$Q5=1.539+1.258\,IP$

図2の一番上にある破線はQ5，その下の実線はF5，1点鎖線はLPを，傾き1の直線はIPをIPで回帰したものである．IPのどの定義域でも，IP<LP<F5<Q5の順に誤分類数が大きくなっている．すなわち，IP-OLDFが一番成績がよく，2次判別関数が一番悪かった．これは，異なった判別手法の新しい評価法になる．

また，これらの誤分類数の平均値に差があることが，t検定で確認できた．

6.2 判別係数の検討

図3は，IP-OLDF，LP線形判別関数，Fisherの線形判別関数の判別係数である．最初の3行は，X4を説明変数とする判別モデルを意味している．1行目と4列目のX4とクロスする−0.556は，IP-OLDFのX4の判別係数である．定数項は全て1なので表から省いている．2行目の−0.625は，LP線形判別関数のX4の判別係数である．3行目の−0.06は，Fisherの線形判別関数の判別係数である．

網掛けしてあるモデルは，注目に値する．X3を説明変数とするモデルでは，IP-OLDFとLP線形判別関数のX3の判別係数が−0.204で同じであり，

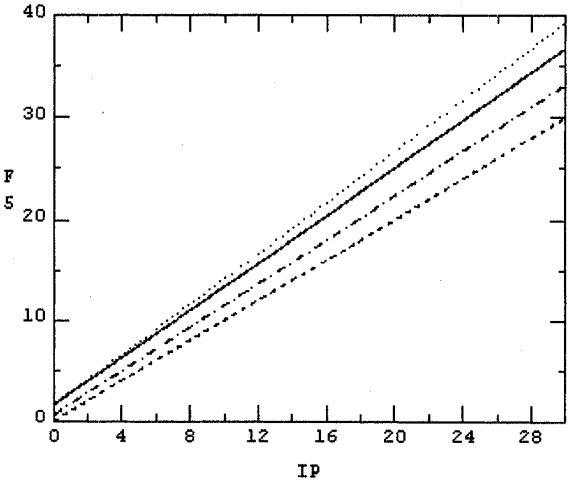

図2　4判別手法の誤分類数のIP-OLDFによる回帰式

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

Fisherの線形判別関数の判別係数は−0.020である．このモデルにX2が追加された（X2，X3）では，IP-OLDFとLP線形判別関数のX2の判別係数が0であり，X3の判別係数が−0.204と同じである．すなわち，モデル（X3）にX2を加えたモデル（X2，X3）において，X2は何ら情報を加えていないことが分かる．これに対して，Fisherの線形判別関数では，X3の判別係数が−0.020なのに対して，モデル（X2，X3）の判別係数は（0.011，−0.027）である．Fisherの線形判別関数からは，説明変数X3をX2に付け加えることの無意味さが分からない．

モデル（X1）に変数X2を加えた（X1，X2）の間でも同じことが言える．すなわち，変数X1にX2を加えても判別に効果が無いことが分かる．

同じことであるが，モデル（X2）に，X1あるいはX3を加えたモデル（X1，X2）と（X2，X3）では，X2の判別係数が0になっている．

これらの知見は，従来の統計的なアプローチから得られなかった数理計画法の功績であろう．すなわち，重回帰分析や判別分析の変数選択問題も，係数が0になるかならないかで判断できれば随分すっきりしたものになる．

6.3 判別結果

図4は，X2を横軸にX4を縦軸にして，（X2，X4）のデータを＋(バージニカ）と□（バーシクル）で2群に分けてプロットしている．実線はIP-OLDFの判別関数を表す．点線はLP線形判別関数，一点鎖線はFisherの線形判別関数を表わす．LP線形判別関数が他の2つと異なっていることが分かる．図から分かりにくいが，IP-OLDFとLP線形判別関数は，端点解を求めているので，少なくとも2個のデータを結んだ直線になっている．

図5は，（X3，X4）を表す．3つの判別関数が異なっていることが分かる．ただし，IP-OLDFの傾きは他の2つの間にある．LP線形判別関数は，2個のバージニカを結ぶ直線であり，2個の＋と□が誤分類されている．それが実線で表わされるIP-OLDFでは2個の＋が誤分類されなくなっている．

図6は，（X1，X3）を表わす．IP-OLDFとLP

	X1	X2	X3	X4
X4				-0.556
				-0.625
				-0.060
X3			-0.204	
			-0.204	
			-0.020	
X1	-0.159			
	-0.159			
	-0.016			
X2		-0.333		
		-0.345		
		-0.035		
X2X4		0.357		-1.190
		-0.062		-0.497
		0.041		-0.129
X3X4			-0.127	-0.253
			-0.152	-0.152
			-0.009	-0.032
X1X3	0.119		-0.357	
	0.092		-0.323	
	0.031		-0.061	
X1X4	0.028			-0.669
	-0.092			-0.253
	-0.000			-0.059
X2X3		0	-0.204	
		0	-0.204	
		0.011	-0.027	
X1X2	-0.159	0		
	-0.159	0		
	-0.015	-0.003		
X2X3X4		0.181	-0.141	-0.482
		0.185	-0.185	-0.370
		0.031	-0.017	-0.062
X1X3X4	0.067		-0.239	-0.146
	0.066		-0.222	-0.192
	0.022		-0.034	-0.042
X1X2X4	-0.000	0.357		-1.190
	-0.108	0.108		-0.394
	-0.006	0.036		-0.098
X1X2X3	0.137	0.103	-0.445	
	0.090	-0.009	-0.315	
	0.033	0.008	-0.067	
X1-X4	0.038	0.070	-0.208	-0.251
	0.060	0.045	-0.232	-0.212
	0.021	0.034	-0.042	-0.074

図3　IP，LPとFisherの判別係数

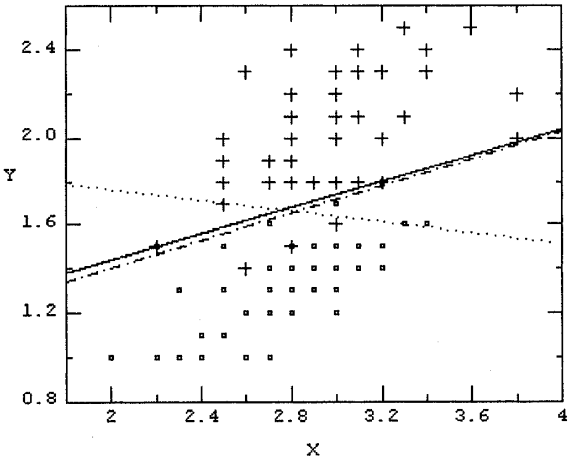

図4　2群のプロット（X2：横軸 vs. X4：縦軸）

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

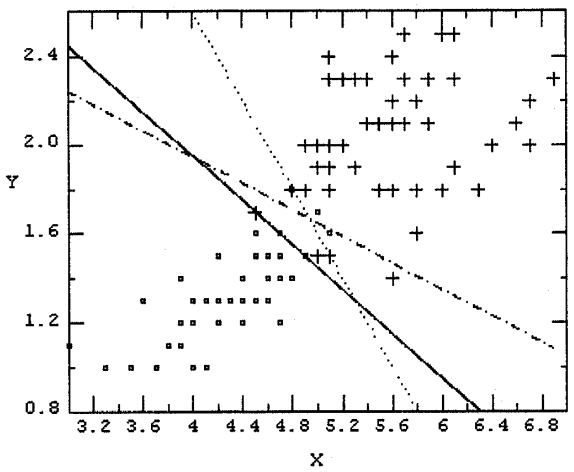

図5　2群のプロット（X3：横軸 vs. X4：縦軸）

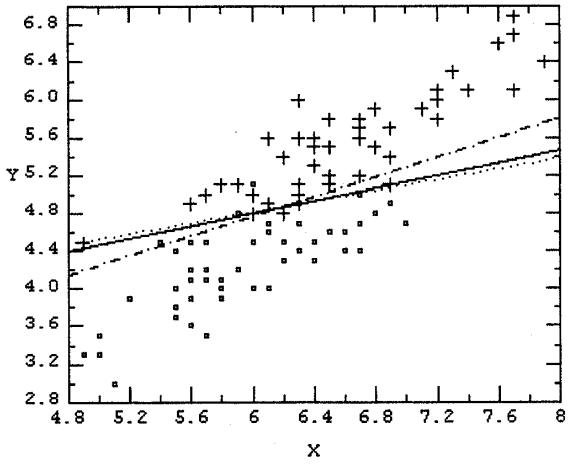

図6　2群のプロット（X1：横軸 vs. X3：縦軸）

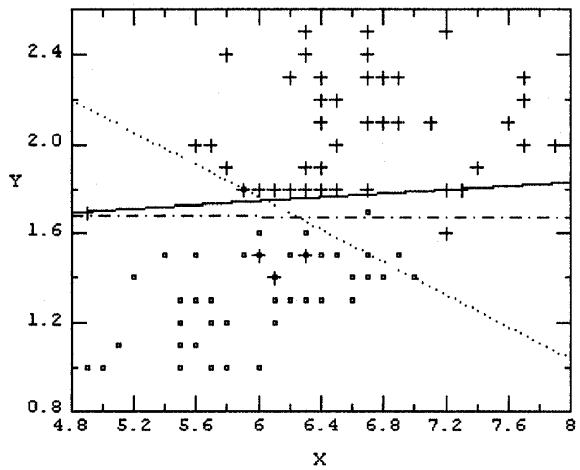

図7　2群のプロット（X1：横軸 vs. X4：縦軸）

線形判別がほぼ重なっているが，Fisherの線形判別関数ともそれ程異なっていない．

図7は，IP-OLDFとFisherの線形判別関数がほぼ同じ傾きであるのに対して，LP線形判別は負の傾きである．Fisherの線形判別関数のX1の判別係数がほぼ0（−0.0002）である．

図8，9は，判別係数で検討したことを表わしている．図8では，X3にX2を加えても何ら判別に貢献しないことを表わしている．すなわち，モデル（X2, X3）はモデル（X3）と本質的に差がないことを表わしている．図9は，X1にX2を加えても意味のないことを表わしている．

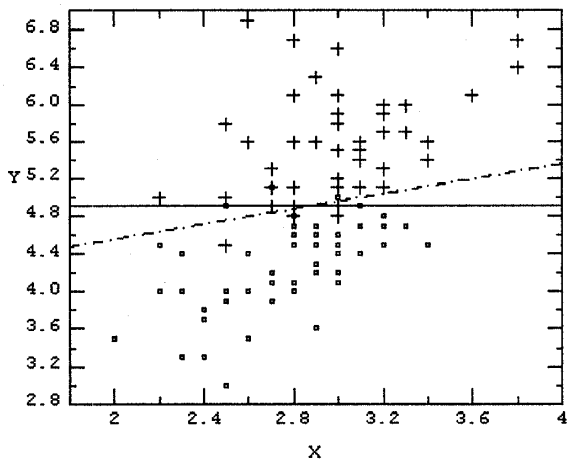

図8　2群のプロット（X2：横軸 vs. X3：縦軸）

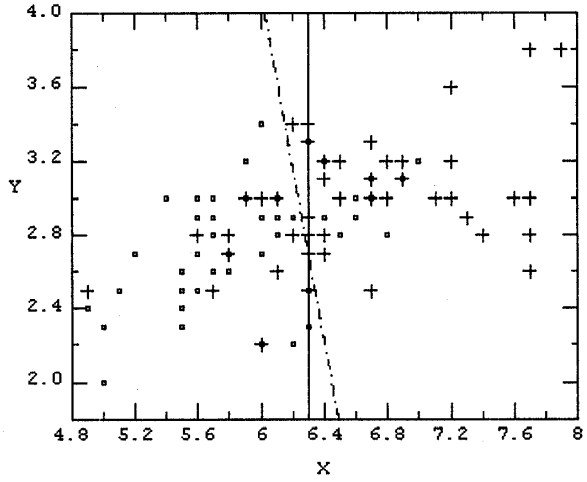

図9　2群のプロット（X1：横軸 vs. X2：縦軸）

7.　まとめと今後の課題

今回，数理計画法を用いたIP-OLDFとLP線形判別関数を提案し，アイリスデータに適用し，Fisherの線形判別関数と2次判別関数との比較評価を行った．

アイリスデータは，説明変数が少なく，基本系列上の誤分類数がたかだか7個であるので，大きな優位性は認められなかったが，それでもIP-OLDFの判別成績が一番良かった．

3月号と4月号では19変数の多重共線性のあるデータと乱数データに適用し，従来の判別分析では得られない興味ある結果について述べる．

参考文献

[1] 新村秀一：パソコン楽々統計学，講談社，1997.

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．