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(1)
—理論的な背景—

新村　秀一

本講座の目的

昔より『近くて遠きは男女の仲』といいますが，ORと統計の関係も似たようなものでしょう．私自身，統計ソフトと数理計画法ソフトの普及や研究活動を通して，ORと統計に関与していながら，これらを結びつける機会になかなかめぐり合いませんでした．この理由としては，ORは研究対象をモデルとして扱うのに対して，統計はデータを確率分布としてとらえ理論展開する点に，なかなか相互に理解できない，深い溝があるのではないかと思っています．

今回，整数計画法を用いた新しい線形判別関数(Optimal Linear Discriminant Function，略してOLDF）を開発し，3種類の異なったデータで評価を行い良い結果を得たので，その概要を以下のように紹介する予定です．

本号では，判別分析の概略とOLDFの定式化を紹介します．

2月号では，Fisherのアイリスデータによる評価を行います[14]．このデータは，判別分析やクラスター分析の評価に用いられ，世界中の統計家が知識を共有する共有財であります．このデータで，OLDFの理解が深まることを意図しています．また，このデータを用いた統計の教科書を作成していますので参考にしてください[13]．

3月号では，多重共線性のある医学データによる評価を行います[14]．回帰分析では，説明変数間に高い相関があれば予測結果が不安定になることは良く知られています．これが多重共線性と呼ばれている悩ましい問題です．判別分析においても同じことがいえます．文献[7，12]では，このデータを用いて多重共線関係のある説明変数の検出，その解消法とモデル決定の方法が提案されています．本講座では，このデータを用いて，2次判別関数やL1ノルム型の判別関数より，OLDFが優れている点，あるいはモデル評価に用いる決定係数の問題点などを指摘します．

4月号では，115組の2変量正規乱数データによる評価を行います[15]．このデータは，現実のデータでは行いにくいクロスバリデーションを目的としております．クロスバリデーションとは，内部標本と呼ばれるデータでモデルを作成し，その結果を評価します．これをInternal Checkと呼びます．実際のデータを分析する場合，この成績だけが示されることが多いのですが，モデル構築に用いていないデータ（外部標本）でExternal Checkをする必要があります．これを行うのは，Internal Checkの結果は内部標本のデータが少ないほど，また用いる説明変数が多いほど，過大に評価（見かけ上，良い結果が得られる）されるからです[1]．クロスバリデーションを行った後，得られたモデルを現実の問題に適用すべきでしょう．最近問題になっている介護保険の評価システムは，内部標本での結果が良かったので，クロスバリデーションを十分行わず，システム化したことに問題があるようです．統計学の社会への大きな貢献の事例になったかもしれないのに，逆の結果になり非常に残念なこです．クロスバリデーションは，応用統計家にとって，決しておろそかにしていけないプリンシプルであります．

一方，理論家は従来新しい理論を開発した場合，その評価を他人任せにしてきました．しかし，PCの能力が飛躍的に向上している今，提案する新手法は提案者自らが評価すべきであると考えています．例えば，2次判別関数は理論の上でも重要であり，興味ある認識を我々に提供しています．すなわち，判別される2群が多次元正規分布で等分散であればFisherの線形判別関数を，そうでなければ2次判別関数を用いるべきということが，これまでの常識でした．しかし，医学データという現実のデータと乱数データで，この指

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

針が正しくない多くの事例が示されています．

これが，今回3種類のデータで評価を行った成果であります．

5月号では，従来の推測統計学に対して深刻な問題を提起しているデータマイニングの中核的な手法である決定木分析による判別結果を紹介する予定でいます．すなわち，これまでの統計学は小標本を想定し，確率分布から帰無仮説が正しいか否かを有意確率（p値）でもって判断していました．しかし，データが増えていけばp値は小さくなっていき，やがて棄却される運命にあります．すなわち，検定論はデータを多く集めれば不用になってしまうわけです．この点で，データマイニングは従来の推測統計学に対して深刻な問題を提示しているわけであります．ORにとっての黒船はいつどこから現れるのでしょうか．

一方，データマイニングも，かつてのAIブームのように熱狂に包まれ，冷静さを欠く面があるようです．今回は決定木分析でIP-OLDFの結果を比較することを目的としていますが，逆に決定木分析の各手法を伝統的な判別分析と比較評価することにもなります．

1. 判別分析について

1.1 判別分析の分類

判別分析の代表的な手法として，Fisherの線形判別関数と2次判別関数がある．これらはパラメトリックな手法で，判別される2群が多次元正規分布であると仮定し，説明変数の1次式と2次式で表される．2群の分散共分散行列が等しければ線形判別関数を，等しくなければ2次判別関数を用いればよいとされてきた．

これに対し，ノンパラメトリックな手法として，離散変量に対してはベイズの定理による判別手法があり，連続変量に対しては最近隣ルールによる手法（複数の群からのマハラノビウスの距離を求め，一番小さな距離を持つ群に判別する．多群判別にも適用可能）がある．

また，分散比，尤度比，誤分類数（あるいは正しく分類された数）という，判定に用いる基準によって分けることができる．分散比による手法としては，Fisherの線形判別関数，正準判別関数，林の数量化II類がある．Fisherの線形判別関数は，2群が多次元正規分布に従い分散共分散が等しいことを仮定している．尤度比による手法としては，多重ロジスティックモデルがある．本講座で紹介するOLDFは，誤分類数を基準とする手法である．この基準は，確率分布を考えないので，多くの統計家には受け入れ難い基準である．すなわち，これまでは標本誤分類数はバイアスに弱く，判別境界点が変わるにつれ変化するので使えないと考えられてきた．しかし，内部標本の誤分類数を直接最小化する基準を取れば，計算時間がかかることを別として，判別分析の理論に貢献できることが分かった．

ベイズの定理とOLDF以外の手法は，最近の統計パッケージに含まれている．これに対して，複数の線形式でデータ空間を分割する区分判別関数のように，提案されたが現実に応用されていない判別手法も数多くある．これは，理論の提案者が，自分自身でその有用性を実証しないで他人任せにしてきた結果であろう．今後は，新しい理論は簡単に検証できるので，その有用性も示し，多くのユーザーを得るように努める必要がある．

しかし，従来の推測統計学のアンチテーゼとして脚光を浴びているデータマイニングの手法には，統計パッケージで忘れ去られたベイズの定理による判別手法を含むものもあるようだ．それゆえ，理論家が，役に立とうが立つまいが，色々な理論を開発することは重要である．それほど才能のない私のような後世の人たちの選択肢が広がるということはありがたいことだ．後世に残る理論を開発できる天才の椅子は限られているが，それらを現実の問題に応用し役立てる作業は数多くあるのではなかろうか．豊かな社会になればなるほど，この比重は高まっていくものと考えられる．

本講座では，整数計画法を用いた新しいIP-OLDFと，これまでも提案されているLPを用いたLP線形判別関数（LP-OLDF）を提案し，従来の判別手法と比較評価することにより，これまでと異なった視点で判別分析を見直したい．

1.2 Fisherの線形判別関数と2次判別関数

(1) Fisherの線形判別関数

Fisherの線形判別関数は，p個の説明変数をもつ2群が，次式で表わされるp変量多次元正規分布に従う確率密度関数をもつと仮定して導出できる．

$$f_i(x)=\{1/SQRT\{(2\pi)^p * |\Sigma_i|\}\} * e^{\wedge}[-(x-m_i)' * \Sigma_i^{-1}(x-m_i)/2]$$

x：説明変数のp次元行ベクトル

m_i：i群のxの平均行ベクトル

Σ_i：i群の分散共分散行列　　(i=1, 2)

2群の分散共分散行列が等しい（$\Sigma_1=\Sigma_2=\Sigma$）として，次の尤度比$f_1(X)/f_2(X)$の対数をとった関数F(x)を

考える．

$$F(x)=\log[f_1(X)/f_2(X)]$$
$$=\{x-(m_1+m_2)/2\}'\Sigma^{-1}(m_1-m_2)$$
$$=x'\Sigma^{-1}(m_1-m_2)-(m_1+m_2)'\times\Sigma^{-1}(m_1-m_2)/2$$

この$F(x)$は，$F(x)=x'\beta+\beta_0$というxの線形関数になっている．これをFisherの線形判別関数と呼ぶ．

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$1/n_1$（あるいは0），2群に$-1/n_2$（あるいは1）を与える．このようにして判別分析のデータを回帰分析のソフトウエアで解析すれば，回帰分析によって得られる回帰係数は，Fisherの線形判別関数の判別係数と比例関係になる．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．

これにより，回帰分析で開発された逐次変数選択法に代表されるモデル選択の知識や統計ソフトが利用できる．また，統計の専門家でない場合，判別分析の理論を特別に学習しないで回帰分析で代用できる．本講座では触れないが，データを工夫することで数量化I類や数量化II類も回帰分析や判別分析で代用できる[9]．このような主張は，林の数量化理論を矯めにする意見ではなく，ORワーカーとして現実問題に対応する場合，思考の節約ができることと，鳥のように広く対象を俯瞰できる利点からの指摘である．

(2) 2次判別関数

2群の分散共分散行列が等しくない場合は，$F(x)$は次のようなxの2次形式になり，2次判別関数と呼ばれる．

$$F(x)=\log[f_1(X)/f_2(X)]$$
$$=x'(\Sigma_2^{-1}-\Sigma_1^{-1})x/2$$
$$+(m_1'\Sigma_1^{-1}-m_2'\Sigma_2^{-1})x$$
$$+(m_2'\Sigma_2^{-1}m_2-m_1'\Sigma_1^{-1}m_1)/2+c$$
$$c:\log[|\Sigma_2|/|\Sigma_1|]$$

線形判別関数に比べて，2次の係数が$p(p-1)/2$だけ増える．このため，内部標本での見かけ上の誤判別率は一般的に良くなるのは当たり前のことである．

2群の分散共分散行列が等しい（$|\Sigma_1|=|\Sigma_2|$）という帰無仮説に対し，χ^2検定で棄却されれば2次判別関数を採用し，棄却されなければ線形判別関数を選ぶことが理論的に推奨されてきた．

本講座では，この頭の中で考え出された指針が，実際のデータへ適用して，妥当か否かも示したい．

1.3 判別関数の問題点

判別関数の問題点は色々あるが，本講座では内部標本と母集団（あるいは外部標本）の関係，事前確率とリスクの扱い，モデル選択の方法論の3点に関して示したい．

(1) 内部標本と外部標本の関係

① 母マハラノビスの距離と標本マハラノビスの距離

各群は多次元正規分布であると仮定する．そして，各ケースxと各群の平均m_iとのマハラノビス（Mahalanobis）の距離MD^2は，次の式で定義される．

$$MD^2=(x-m_i)'\Sigma_i^{-1}(x-m_i)$$

そして，マハラノビスの距離MD^2の大小で，どの群に属するかが決定される．すなわちマハラノビスの距離は，判別空間における2つのデータ間の距離を表している．マハラノビスの距離が大きいほど離れており，小さいほど近いことになる．この距離はユークリド距離の上に，データの確率密度を考慮したものと理解しておけばよいだろう．

このとき，2群の母マハラノビスの距離δと標本マハラノビスの距離Dとの関係は，次のように要約される[1]．

標本マハラノビスの距離は，標本数が少ないほど，あるいは変数が多いほど，母マハラノビスの距離より大きくなる確率が高くなる．すなわち，少ないデータや，多くの変数を用いて判別分析を行うと，標本誤判別率$\Phi(-D/2)$は母誤判別率$\Phi(-\delta/2)$より過小評価される確率が高くなる．ただし，Φは規準正規分布関数である．

すなわち，少ないデータで多くの説明変数を用いて分析すれば，判別成績がよくなって当たり前で，それを現実のデータに適用すれば惨憺たる結果になることを意味する．あるいはシミュレーションなどで，制御可能な変数を数多く用いてモデル化すれば，作成者の意図のままどのような結果を得ることも可能であることと対応している．

このため，判別結果がOverestimateされない良いモデルを作成するためには，モデル構築に用いる質の高いデータをできるだけ多く集めるか，モデル作成に用いる変数はできるだけ少ない方がよいという「ケチの原理（Principal of parsimony）」あるいは「オッカムの剃刀」の警告に従う必要がある．このため，回帰分析では，変数を少なくするために逐次変数選択法などの手法が提案されてきた．本講座では，従来の変

数選択法の手法で用いられてきた逐次F検定といわれるまったく無意味な停止則に疑問を投げかけ，その理論的な意義は認めつつ，停止則を用いない基本系列で比較評価することの有用性を医学データで示す．また，モデルの良さを示す決定係数のような従来の統計量に対して，最小誤分類数の有用性を紹介する．

② External Check

一方，母集団と標本の関係を実際に比較することはできないので，それに代わって外部標本によるExternal Checkや，Jack knife法が用いられる．External Checkは，内部標本で得られた判別関数を，外部標本に適用して評価することである．すなわち，母集団と標本集団の関係は，「少ないデータや，多くの変数を用いて得られた判別関数は，外部標本に適用すると一般的に悪くなる」と読み替えることができる．すなわち，内部標本で誤判別率が良くても，外部標本の誤判別率が悪ければ，その判別関数を採用しないことになる．

データが少ない場合，それらを内部標本と外部標本に分割することは心理的に抵抗が大きいので，Jack knife法などの代替法が用いられる．

③ 問題点

この点から，同じデータに対してFisherの線形判別関数と2次判別関数を適用しても，後者は推定するパラメータ数が増えるので，見かけの誤判別率は良くなって当たり前である．しかし，2次判別関数の理論は精緻であり魅力的なので，あまり批判は行われてこなかった．

一方，内部標本の誤分類数を最小化する基準で導出されるOLDFは，以上の議論とは異なるが，過度に内部標本に適合し外部標本に対しては悪くなるのではないかという危惧がある．すなわち，特定のデータに依存し，結果が不安定であったり，普遍的な結果と縁遠いと考えられる．

本講座では，以上の2点に関して具体的なデータの分析結果で常識を覆す結果が得られることを明らかにしたい．

(2) 事前確率とリスク

① 尤度比

Fisherの線形判別関数では，尤度比(f_1/f_2)が1すなわち2群の確率密度が等しい点を判別境界点と呼んで，判別している．これを，尤度比方式による判別という．この場合$F(x)=0$は，p次元のデータ空間を2分する超平面になる．そして，データxを代入して$F(x)>0$であればG1群（あるいはG2群）と判別し，$F(x)<0$であればG2群（あるいはG1群）と判別することになる．

② 事前確率とリスク

しかし，2群の母集団のサンプルサイズが異なる場合，その違いを事前確率π_1，π_2で表し，尤度比として$(\pi_1 f_1)/(\pi_2 f_2)$を考える．すなわち，判別境界点は0の代わりに，$\log(\pi_2/\pi_1)$になる．

さらに，医療診断で正常群と疾病群（企業診断では，優良企業と倒産企業と読み替える）を考えた場合，疾病群の事前確率が小さくても誤分類によるリスクは正常群より大きく考える必要がある．この場合，リスクをr_1，r_2とすると，尤度比は$(r_1\pi_1 f_1)/(r_2\pi_2 f_2)$になる．結局のところ，事前確率やリスクを考えた場合の判別境界は0でなく，$\log\{(r_2\pi_2)/(r_1\pi_1)\}$になる．

以上の通り，事前確率とリスクは現実問題においては重要であるが，その値を決定することは難しい面もある．また，判別分析の評価のために種々の統計量が提案されているが，医学や企業などの実務分野では分かりやすい誤分類数がよく用いられている．しかしすでに述べた通り，事前確率とリスクの違いで，誤分類数が異なるという問題点がある．

③ ROC曲線

そこで，事前確率とリスクの違いは，$F(x)$の定数項の違いで表されることを利用して，それらの判別境界点を何段階かで変えて，2群の正診率（True Positive）と誤判別率（False Positive）をx-y平面にプロットしたROC曲線（Receiver Operating Characteristic Curve）を描いて，各種判別結果の優劣の比較評価を行うことが考えられる[9]．

例えば，文献[5]では，胃X線像のデータを各種判別分析（枝分かれ法，数量化II類，ベイズ診断，多重ロジスティック，主成分分析，判別分析）で分析した結果と医師診断の結果をROC曲線で比較している．

(3) モデルの評価

IP-OLDFの誤分類数は標本に対して一意に決まり，変数を増加すると単調に減少するという性質がある．

本講座では，IP-OLDFの誤分類数でもって，他の判別関数の誤分類数を評価することを勧めたい．これによって，従来評価の難しかった判別手法の優劣を定量化できる．

2. OLDFのアルゴリズム

2.1 探索的なOLDFとIP-OLDF

文献[2～4]で，内部標本の誤分類数を最小化する探索的なOLDFが提案されている．現実のデータにおいて，多次元正規分布を仮定することには多くの場合無理があり，特定の分布を仮定しないところに本手法の意義がある．すでに述べたとおり，内部標本の誤分類数を最小化するということは，内部標本に過敏にフィットし，外部標本の誤判別率は逆に悪くなる恐れがある．

一方，OLDFによる内部標本の誤判別率がFisherの線形判別関数に比べて少なければ少ないほど，対象データが正規性から乖離していることを分り易く示してくれるのではないかと期待できる．また，OLDFの誤分類数は内部標本に対し一意に決まるので，事前確率やリスクの導入によって誤分類数が異なり判別手法の比較評価が難しいという点を解決できる．

本講座では，探索的なOLDFに代わって，新しく整数計画法を用いて定式化されたIP-OLDFを勧めたい．探索法に代わって，数理計画法を用いることで真の最適解が求まり，将来的には数理計画法の有用な情報から確率論とは別の組み合わせ論による判別分析の知見が得られるかもしれない．また，判別分析と回帰分析などを含む線形モデルとして広い枠組みで考察できる．

2.2 探索的なOLDFのアルゴリズム

(1) p次元データ空間における判別

線形判別関数の定数項を1に規準化した線形判別関数 $(f(x)=bx'+1)$ を考える．

G1群に属するデータ x_i が，$f(x_i)>0$ ならば正しく判別され，$f(x_i)<0$ ならば誤分類されたとする．この場合，G2群に属するデータ x_i では，$f(x_i)<0$ ならば正しく判別され，$f(x_i)>0$ ならば誤分類されることになる．

しかし，G2群の不等式の両辺に－をかけることで，$-f(x_i)>0$ ならば正しく判別され，$-f(x_i)<0$ ならば誤分類されることになる．こうすることで，不等号の向きをG1群の場合と同じに統一できる．

このような統一された表記法において，判別スコアが負になるものを数え上げれば，それが内部標本の誤分類数になる．

(2) 判別係数の空間における判別

次に，個々のデータ x_i を線形判別関数の係数とする線形式 $Hi\ (g_i(a)=x_ia'+1)$ を考える．すなわち，判別係数の空間において個々のデータは，判別係数の空間を2分する超平面になる．データ空間を判別係数の空間に置き換えて考えることが，探索的なアルゴリズムの重要な点である．

G1群に属するデータ x_i で，$g_i(a)>0$ になる半平面を線形式Hiの＋半平面とし，$g_i(a)<0$ になる半平面を線形式Hiの－半平面とする．線形式Hiの＋半平面に含まれる任意の係数aでデータ x_i は正しく判別され，－半平面の任意の係数aでデータ x_i は誤分類されることになる．

G2群のデータ x_i では，両辺に－をかけることにより，$-g_i(a)>0$ を＋半平面とし，$-g_i(a)<0$ を－半平面とすることで，G1群と同じ不等号の向きになる．

判別係数の空間aは，このようにしてn個のデータ x_i で作られるHiで2分割され，その結果として凸体に分割される．この凸体の内部の点で，全ての線形式Hiの＋半平面に含まれる個数を数え上げる．それは，その凸体の内部の点を判別係数とする判別式において，正しく判別されたデータの個数になる．一方，－半平面に含まれる個数を数え上げれば誤分類数になる．すなわち，凸体の内部の点に対して，正しく判別された個数（あるいは誤分類数）が一意に決まる．

このようにして，p次元判別係数の空間は，Hiによって有限個の凸体に分割され，凸体の内部の点（判別係数）は，同じ誤分類数（あるいは正しく判別された数）をもつことになる．

(3) 簡単な例

例えば，図1のようにデータが3件ある判別の問題を考える．1件目のデータがG1群に属し，$(-0.14, -0.21)$ という値を取るとすれば，これを係数とした判別関数は $-0.14a-0.21b+1$ で表される．線形式 $-0.14a-0.21b+1=0$ をH1とする．H1によって，判別係数の空間は2分割される．$-0.14a-0.21b$

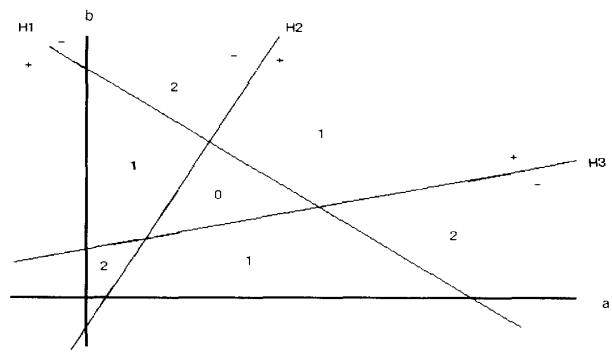

図1 Example of OLDF

$+1>0$ をH1の＋半平面と呼ぶ．この半平面の任意の点(a, b)は，データ空間の判別係数に対応する．$-0.14a-0.21b+1>0$ であるから，データ空間で考えた判別関数 $ax+by+1>0$ は，データ $(-0.14, -0.21)$ を正しく判別する．すなわち，H1の＋半平面にある点（判別係数）は，H1に対応したデータを正しく判別する．一方，$-0.14a-0.21b+1<0$ をH1の－半平面と呼ぶ．この空間の任意の点(a, b)は，$-0.14a-0.21b+1<0$ であるから，データ空間で考えた判別関数は $a*(-0.14)+b*(-0.21)+1<0$ になりデータ $(-0.14, -0.21)$ を誤判別する．すなわち，H1の－半平面にある点（判別係数）は，H1に対応したデータを誤判別する．

2件目のデータはG2群に属し$(-2, 2)$としよう．線形式H2は $-2a+2b+1=0$ であるが，G2群に属するので両辺に－をかけた $2a-2b-1=0$ をH2とする．$2a-2b-1>0$ がH2の＋半平面であり，$2a-2b-1<0$ が－半平面になる．

すなわち，図1は3つのデータで表される線形式{H1, H2, H3}の場合を考えている．これらの線形式で，2次元の平面は，7個の凸体に分割される．真中にある三角形の内部の点は，3つの線形式で作られた＋半平面の側にあるので，この内部の点で表される判別係数をもつ判別関数は，3個のデータを正しく分類し，誤分類数はゼロになる．この三角形と辺を共有する3つの凸体は，その辺すなわち線形式のお互い反対側にあるので，三角形側が＋半平面であれば，反対側の凸体の誤分類数は1増える．逆に，もし三角形側が－半平面の側にあれば，反対側の誤分類数は1少なくなる．このようにして，辺を共有する凸体は，誤分類数がお互いに1異なることになる．

ここで重要なことは，最適凸体は必ず＋半平面で囲まれていることである．もし－半平面のものがあれば，反対側の凸体の誤分類数が1少なくなり，現在の凸体が最適でなくなることになる．

(4) 探索的OLDFのアルゴリズム

三宅と新村[4]は，p次元空間で誤分類数を最小にする凸体を探索的に探すアルゴリズムを提案している．

簡単に紹介すれば，最初はa軸とb軸を初期座標とする．a軸とは，H3, H2, H1の順に交差している．交差点に接する凸体の誤分類数を調べ上げると，H2とH1の交差点に接する凸体（今の場合は同じもの）の誤分類数が1と一番小さい．次に，b軸と交差する点に接する凸体で最小誤分類数を持つものを調べる．この場合も，誤分類数が1の2つの凸体が選ばれる．次に座標軸を変更する．候補は，{a, H2}, {a, H1}, {b, H2}, {b, H3}, {b, H1}である．これらの新座標系で，それぞれ探索し，誤分類数最小のものを探索する．座標変換で最小誤分類数を更新できなければ，得られた凸体の頂点を原点とする座標系で探索する．図1の場合，{H1, H2}, {H3, H1}, {H2, H3}である．この探索で，最小誤分類数が更新されなければ，この凸体を一応最適凸体として終了する．OLDFの判別係数は，最適凸体の内部のどの点でもよいが，一応全ての頂点の単純平均値を用いることが考えられる．

この方法は，残念ながらその当時のIBM汎用機を用いても計算時間がかかり，19変数の今回用いている医学データでは6変数までしか分析できなかった[4]．また，探索的OLDFのアルゴリズムは得られた解が大域的な最適解か否かの保証がない欠点がある．

本講座は，20年後に著しく計算速度が改善された整数計画法を用いた新しい最適化手法の紹介である．

2.3 IP-OLDFのアルゴリズム

(1) 数理計画法の回帰分析への応用

数理計画法の回帰分析への応用として，重回帰分析が2次計画法（Quadratic Programming, QP）で，LAV（Least Absolute Value）回帰分析が線形計画法（Linear Programming, LP）で定式化できることは古くから指摘されている[8, 11]．さらに一般化して，非線形計画法でLpノルムの回帰分析を扱うこともできる．

しかし，最小自乗法による回帰分析を，わざわざ2次計画法で行うことは，操作性や計算時間の点からも得策ではない．これに対して，LPを用いて誤差の絶対値の和を最小にするLAV回帰分析は，すでにSASなどの商用統計ソフトにも提供されている[6]．

(2) 数理計画法の判別分析への応用

すでに述べた通り，線形判別関数は回帰分析の特殊例であった．線形計画法や2次計画法による判別関数の定式化は，その延長線上で行える．

Glover[10]は，線形計画法を用いて定式化された判別関数のそれまでの研究を集大成し，それらを含む改良モデルを提案している．スラック変数とサープラス変数を個別データ毎に導入し，それらに対して重みを導入することで，誤分類されたケースの判別境界点からの距離の加重和から正しく分類されたケースの判別境界点からの距離の加重和を引いたものを目的関数

としている．そして，それを最小化する線形計画法のモデルとして定義している．誤分類されたケースの重みを1とし，正しく分類されたケースの重みを0としたものが，本講座で取り上げるLP線形判別関数である．

このように重みを適当に与えることで，LPで定式化できる判別関数のモデルを包括的に記述した点で，Gloverの研究は評価できる．

しかし，実際のデータへの適用にはいたっておらず，小さな例題で検証しているのみである．また，せっかく最適化手法を用いているのに，恣意的な重みを用いることで，得られた結果の解釈が難しくなる問題もある．

これに対して，整数計画法を用いた判別関数の研究は一般的でない．整数計画法のような組み合わせ最適化は，計算時間がかかり，欧米の大学で広く利用されているシカゴ大学で開発されたLINDOの普及を行ってきた筆者の経験では，実用問題に適用できるようになったのは1990年代の中頃に入ってからである．そのため，本来IPで定式化すべき問題もLPで定式化することが，これまでは数理計画法でのモデル作成のひとつの見識であった．

(3) IP-OLDFの定式化

整数計画法を用いた最適判別関数は，定式化すると次のようになる．

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

ST

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

・

・

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

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

・

・

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

END

最初のm個の制約式はG1群のデータに対応した制約式であり，その後のn個の制約式はG2群のデータに対応している．cは大きな正の定数であり，e_iは各データx_iに対応し，各データが誤分類されれば1に，正しく分類されれば0になる0/1型の整数変数とする．こうすることで，正しく分類されるデータの判別境界点は0のまま，誤分類されるデータの判別境界を0から$-ce_i$に緩めることができる．

目的関数は，e_iの和すなわち判別境界点を緩める制約式の数（内部標本の誤分類数）を最小化している．

そして，これでもって内部標本の誤分類数を最小にする凸体の頂点（判別係数a）が最適解として求められる．ただし，数理計画法モデルの特徴として，個別データ毎に重みづけすることは容易で，各e_iに重みを掛ければすむことである．こうすることで，Gloverのように，色々な派生手法を新たに定義できる．しかし，恣意的な重みの利用は，最適化のメリットを相殺すると考える．

実際には，不等号を逆にしたものも計算し，誤分類数の少ないほうを選ぶ必要がある．また，得られた解は最適凸体の頂点である．最終的には，探索的OLDFで紹介したように全ての頂点を求め単純平均で得られた値を判別係数とする必要がある．

(4) LP線形判別関数の定式化

一方，c=1として，e_iを正の実数とすれば，IP-OLDFのモデルを，計算時間のかからないLPによる線形判別関数として定式化できる．LP線形判別関数は，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．これに対し，IP-OLDFは，誤分類されるケースの個数の和を最小化している．すなわち，誤分類されるケースは同じウエイトになる．

Gloverは，次のLPモデルで判別関数を定式化している．

$$\text{MIN} \quad h_0u_0+\Sigma h_iu_i-k_0v_0-\Sigma k_iv_i$$

ST

$$a_0+x_{1j}a_1+\cdots+x_{pj}a_p+u_0+u_j-v_0-v_j=b$$

(ただし，j=1, …, n)

END

すなわち，u_jは誤分類されたデータの判別境界点からの距離である．u_0はデータ全体の平均である．一方，v_jは正しく分類されたデータの判別境界点からの距離である．v_0はデータ全体の平均である．目的関数は，これらの距離に重みを掛けて，誤分類された距離の和から正しく分類された距離の和を引いて，最小化することを提案している．これにより，誤分類された距離の和を最小化し，正しく分類されたデータの距離の和を最大化するという，多目的最適化をLPで定式化したことになる．

LP線形判別関数は，$h_0=h_i=1$，$k_0=k_i=0$にしているので，Gloverのモデルの特殊例になる．

このように個別データ毎に重みu_jとv_jを与え多様

なモデルを定義できるが，恣意的な判断が必要となり，その導入は慎重を要する．

また，IP-OLDFとの比較のため，今回の定式化を採用する．別の機会に，Gloverのモデルで再評価してみる必要があるだろう．

以下次号から，IP-OLDFとLP線形判別関数を，3種類のデータを用いて既存の2つの判別関数と比較評価する．すなわち，新しい手法は既存の手法と複数のデータで評価することが重要である．

本講座の多くは，成蹊大学研究助成金に負っている．

参考文献

[1] A. Miyake & S. Shinmura, Error rate of linear discriminant function, F. T. de Dombal & F. Gremy, editors 435-445, North-Holland Publishing Company, 1976.

[2] A. Miyake & S. Shinmura, An Algorithm for the Optimal Linear Discriminant Function, Proceedings of the International Conference on Cybernetics and Society, 1447-1450, 1978.

[3] S. Shinmura & A. Miyake, Optimal linear discriminant functions and their application, Proceedings of the COMPSAC 79, 167-172, 1979.

[4] 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.

[5] 新村秀一，鈴木隆一朗，中西克巳：胃X線像の各種判別分析，オペレーションズ・リサーチ，26-1, 51-60, 1981.

[6] J. P. Sall, SAS Regression Applications, SAS Institute Inc, 1981[新村秀一 (1986), SASによる回帰分析の実践，朝倉書店].

[7] 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.

[8] 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.

[9] 森村英典，牧野都治編：統計・OR活用事典，東京書籍，1984.

[10] F. Glover, Improve Linear programming models for discriminant analysis, Decision Sciences, 2, 771-785, 1990.

[11] 新村秀一，高森寛：実践数理計画法—LINDOを用いて—，朝倉書店，1992.

[12] 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—，成蹊大学経済論集，27-1, 180-203, 1996.

[13] 新村秀一：パソコン楽々統計学，講談社，1997.

[14] 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, 89-101, 1998.

[15] 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, 107-124, 1999.

[16] S. Shinmura, A new algorithm of the linear discriminant functions using integer programming, New Trends in Probability and Statistics Volumn 5, 133-142, 2000.

[17] S. Shinmura & T. Tarumi, Evaluation of the optimal linear discriminant functions using integer programming (IP-OLDF) by the normal random data, Proceedings in Computational Statistics 2000, 95-96, 2000.