A－09：K-重交差検証法による改定 IP-OLDF と S-SVM, LDF，ロジスティック回帰の評価

成蹊大学　経済学部　新村秀一

0.　研究の動機

1948 年大学を卒業し，大阪府立成人病センターで「心電図の自動診断解析システム」の診断論理を，判別分析で研究したが，医師の枝分かれ論理にかなわなかった．そして，医学診断のように判別超平面上に異常群のケースが多い判別には LDF は適していないと考えた．

1978 年から 1980 年まで，**MNM(最小誤分類数)基準**による最適線形判別関数の研究を行ったが，ヒューリスティック手法のため研究を発展させることができなかった．

1.　MNM 基準による IP-OLDF と改定 IP-OLDF　(1997-2010)

1.1　IP-OLDF と 2 つの新知見

1978 年に SAS を日本に紹介し，1983 年に数理計画法ソフトの LINDO を日本に紹介したことから，混合整数計画法による MNM 基準による **IP-OLDF** を自然に導いた．

2 個の説明変数で 3 ケースを判別する IP-OLDF は次のようになる．

CLASS1：$z_1=(-1/18, -1/12)$，　CLASS2：$z_2 = (-1, 1/2)$，$z_3 =(1/9, -1/3)$

$MAX=\Sigma e_i$；　$-(1/18) * b1 - (1/12) * b2+1 \geqq -e_1$；

$-\{ -b1 + (1/2) * b2 + 1 \} \geqq -e_2$；　$-\{(1/9) * b1 - (1/3) * b2 +1\} \geqq -e_3$；

・定数項を 1 に固定しているので，p 次元のデータ空間と判別係数の空間の両方で解釈でき，次の 2 つの新知見を得た．注：定数項を固定しないと(p+1)次元になる。

新知見 1:　2 変数の値を係数とする 3 個の線形超平面（制約式の等式に対応）は，判別係数の空間を 2 つの半平面に分割．$y_i*(\mathbf{x}_i'\mathbf{b}+1)>0$ の＋半平面に含まれる判別係数 $\mathbf{b}_j$ は $\mathbf{x}_i$ を判別し，$y_i*(\mathbf{x}_i'\mathbf{b}+1)<0$ のー半平面に含まれる判別係数 $\mathbf{b}_j$ は $\mathbf{x}_i$ を誤判別する．3 個の線形超平面は，判別係数の空間を 7 個の凸体に分割する．凸体の内点は，ー半平面の数に対応する同じケース $\mathbf{x}_i$ を誤分類し，凸体の頂点と辺に対応する判別関数は判別超平面上にケースがあり誤分類数の決定が正しく行えないことがある．凸体の内点の誤分類数が最小のものを**最適凸体**と呼び，改定 IP-OLDF はこの内点を求めるが，IP-OLDF はデータが一般位置にない場合，正しい最適凸体の頂点を求めないことがあることが分かった．他の判別関数は凸体の内点を求めることが理論的に保証されないので，得られた誤分類数は正しか正しくないか分からない．

新知見 2:　MNM の単調減少性（$\mathbf{MNM_p \geqq MNM_{(P+1)}}$）

1.2　改定 IP-OLDF と 2 つの判別分析の問題

判別規則($y_i*f(x_i)>0$ で x_i は正しく判別され，$y_i*f(x_i)<0$ で誤判別)の単純さに隠され，判別分析の問題が隠されてきた．　**改定 IP-OLDF は理論的にこの問題を解決．**

問題 1:　判別超平面上（$y_i*f(x_i)=0$）のケース x_i の帰属は一般的に分からない．すなわち判別関数は正しい誤分類数が分からない．ただし，判別規則が説明変数で記述できる場合は，この問題は生じない（試験の合否判定を大問の得点で判別する 3.2 の例を参照）．

問題 2：SVM は判別分析を線形分離可能（MNM=0）な判別を出発点にしたが，MNM=0 の判別分析の問題点が研究されていない．分散共分散行列に基づく Fisher の線形判別関数（LDF）や 2 次判別関数は，MNM=0 のデータを多くの場合に正しく認識できない（誤分類数が大きい）．

問題 3:MNM 基準に基づく改定 IP-OLDF は学習標本で過学習し検証標本で汎化能力が悪く，LDF は正規分布を仮定しているので汎化能力が良いと，事実に基づかないで信じられてきた．しかし，Fisher のアイリスデータ（LDF の評価に利用），スイス銀行紙幣データ（2 変数で MNM=0 の事実が IP-OLDF で初めて指摘），CPD データ（多重共線生），学生データ（一般位置になく，判別超平面上のケースの問題の理解に最適）の 4 種の実データからリサンプリング法を用いて 100 倍の大標本を作成し，100 重交差検証法で改定 IPLP-OLDF(MNM の近似解を高速で求解)で，LDF とロジスティック回帰が学習標本と検証標本で圧倒的に悪いことが分かった．

	LDF-改定 IPLP		ロジスティック回帰 - 改定 IPLP	
手法(**135**)	学習	検証(**15**)	学習(**3**)	検証(**33**)
アイリス(**15**)	[0.55, 5.23]	[-0.6(**2**), 2.36]	[0.59, 5.31]	[-0.84(**2**), 1.85]
銀行(**63**)	[0, 5.32]	[-0.33(**10**), 3.45]	[0, 5.40]	[-0.3(**24**), 3.64]
学生(**31**)	[1.46, 8.61]	[-1, 29(**3**), 7.11]	[-2.12(**3**), 6.48]	[-2.89(**7**), 5.59]
CPD(**26**)	[3.05, 7.28]	[2.21, 6.15]]	[0, 13, 3.43]	[0.29, 1.74]

２．応用研究（2011～2013）

10 択 100 問の試験の大問 4 問の得点を説明変数とし，得点の 10%点，50%点，90%点の 3 水準で 18 個の合否判定の判別を行い LDF や QDF の誤分類確率が悪かった．小標本のための 100 重交差検証法で，改定 IP-OLDF, LDF, ロジスティック回帰，S-SVM の比較を行った．学習標本と検証標本で改定 IP-OLDF が良く，LDF はロジスティック回帰と S-SVM と比べても極端に悪かった．また，2 次判別関数と正則化判別分析で，驚く問題が観測された．