2001年度日本オペレーションズ・リサーチ学会
春季研究発表会

目標計画法を用いた判別分析法

01205520　東京理科大学　末吉　俊幸　SUEYOSHI　Toshiyuki
東京理科大学　＊吉田　実恵子　YOSHIDA　Mieko

1. はじめに

Altman[1]以来、判別分析法(Discriminant Analysis,以下DA)は幅広く適用されてきた。DA は一般的に統計的DA か目標計画法(Goal Programmin, 以下 GP)ベースのDA に分類され、前者の方法的拡張にはロジット・プロビットモデルという手法がある。しかし一般的な統計的DA は、データが多変量正規分布に、ロジット・プロビットモデルは誤差が分布に従うという仮定が必要である。一方 GP ベースの DA は分布に従わない等の統計的性質を持っているにもかかわらず、現実問題に用いられるレベルまで統計的に発展されていなかったため、あまり適用されなかった。そこで本論文では、GP ベースの DA のために 2 つのノンパラメトリック検定の使用を新しく提案し、統計的に試みる。

2. ロジットモデルとプロビットモデル

ロジット・プロビットモデルは、OLS(通常の最小2乗法)では解けないモデルの従属変数の範囲を実数線から 0-1 間に変形する際に使用され、前者はロジスティック関数、後者は累積標準正規分布を用いる。その際潜在する従属変数 z_j が派生し、次の法則が適用される：$y_j = 0 \Leftrightarrow z_j \le \lambda (j \in G_0), y_j = 1 \Leftrightarrow \lambda < z_j (j \in G_1)$。

3. 判別分析(DA)

3.1 L_1回帰モデルのための GP 公式

$$\begin{aligned} \min \quad & \sum_{j \in J} (\delta_j^+ + \delta_j^-) \\ \text{s.t.} \quad & \sum_{j=1}^{m} \beta_i x_{ij} + \delta_j^+ - \delta_j^- = z_j, \ (j \in J \equiv G_0 \cup G_1) \\ & z_j \le \lambda, (j \in G_0) \\ & \lambda + \eta \le z_j, (j \in G_1) \\ & \beta_i, z_j \text{は制約なし} \qquad \delta_j^+, \delta_j^- \ge 0 \end{aligned} \tag{1}$$

(1)式は上記モデルの代替案となり、δ_j^+ は分離関数 $\sum_{i=1}^{m} \beta_i x_{ij}$ から上の垂直分布であり、δ_j^- は下の垂直分布である。η(微小数)は誤判別なし等の自明解を避けるために加えられ、ここでは 1 とする。目的関数は z_j と $\sum_{i=1}^{m} \beta_i x_{ij}$ 間の絶対偏差の和を最小化する。(1)式の最適解 β_i^* と λ^* はパラメータ評価値である。

3.2 GP ベースの DA の計算過程

オーバーラップがある時、(1)式は再公式化される。

Stage 1：オーバーラップの検証

$$\begin{aligned} \min \quad & \sum_{j \in G_0} \delta_{0j}^- + \sum_{j \in G_1} \delta_{1j}^+ \\ \text{s.t.} \quad & \sum_{i=1}^{m} \beta_i x_{ij} + \delta_{0j}^+ - \delta_{0j}^- = \lambda, \qquad (j \in G_0) \\ & \sum_{i=1}^{m} \beta_i x_{ij} + \delta_{1j}^+ - \delta_{1j}^- = \lambda + \eta, \quad (j \in G_1) \\ & \beta_i, \lambda \text{は制限なし} \quad \text{すべての} \delta \text{変数} \ge 0 \end{aligned} \tag{2}$$

全体集合(J)は $J = G_0 \cup G_1 = C_0 \cup C_1 \cup R_0 \cup R_1$ と分ける。下位集合 C_0 は G_0 に、C_1 は G_1 正しく分類された観測値集合を示し、誤分類は R_0 と R_1 で検証される。

Stage 2：オーバーラップの分類に取り組む

$$\begin{aligned} \min \quad & \sum_{j \in R_0} \delta_{0j}^- + \sum_{j \in R_1} \delta_{1j}^+ \\ \text{s.t.} \quad & \sum_{i=1}^{m} \beta_i x_{ij} \le \lambda, \qquad (j \in C_0) \\ & \sum_{i=1}^{m} \beta_i x_{ij} + \delta_{0j}^+ - \delta_{0j}^- = d \quad (j \in R_0) \\ & \sum_{i=1}^{m} \beta_i x_{ij} + \delta_{1j}^+ - \delta_{1j}^- = d \quad (j \in R_1) \\ & \sum_{i=1}^{m} \beta_i x_{ij} \ge \lambda + \eta \qquad (j \in R_1) \\ & \beta i, \lambda, d \text{は制約なし} \quad \text{全ての} \delta \text{変数は} \ge 0 \end{aligned} \tag{3}$$

新しいサンプル(X_s)は次の法則で分類される。

(a) $\sum_{i=1}^{m} \beta_i^* x_{is} < d^*$ の時、X_s は G_0 に分類される。

(b) $\sum_{i=1}^{m} \beta_i^* x_{is} > d^*$ の時、X_s は G_1 に分類される。

(c) $\sum_{i=1}^{m} \beta_i^* x_{is} = d^*$ の時は事前情報、我々の決定による。

© 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．

4. 判別分析法のためのノンパラメトリック検定

4.1 Kruskal-wallis 検定

G_1 と G_2 の分布を統計的に知るため、以下の帰無仮説を立て、Step に沿って Kruskal-wallis 検定を行う。

H_0 ：G_1 と G_2 の観測値の母集団は同じである

Step1：GP ベースの DA を 2 グループの観測値に適用する。Stage2 で、判別得点値(d*)とパラメータ評価値 $\beta_i (i=1,\cdots,m)$ を得る。

Step2： $\sum_{i=1}^{m} \beta_i^* x_{ij}$ (C_j とする)を計算する。C_j の大きいほうから全観測値は rank=1、rank=2 とランク(R_j)付けされる。n_0 は G_0、n_1 は G_1 の観測値数を表し、$n = n_0 + n_1$ である。

Step3： $W_0 = \sum_{j \in G_0} R_j$ 、 $W_1 = \sum_{j \in G_1} R_j$ とした時、統計値 H は(4)式で与えられ、自由度 k-1 で χ^2 分布に従う。

$$H = \frac{12}{n(n+1)}\left(\frac{W_0^2}{n_0} + \frac{W_1^2}{n_1}\right) - 3(n+1) \qquad (4)$$

Step4：$H \geq \chi_\alpha^2$ の場合、帰無仮説 H_0 は有意水準 α% (片側)で棄却され、それ以外では棄却されない。

<同点の観測データの扱い方>

同点な値を持つ集合が 2 つ以上あった場合、統計量 H は(5)式を用いて調整される。

$$1 - \sum \tau / n^3 - n \qquad (5)$$

ここで τ は C が同点の集合内での観測値の数である。統計量 H は(6)式に書き換えられる。

$$H^C = H \Big/ \frac{\sum \tau}{n^3 - n} \qquad (6)$$

4.2 感度分析

次に、新しい帰無仮説 H_0 を立て再び検定する。

H_0 :β_i $(i=1,...,m)$ は判別率に影響しない。

Step 1：前節同様に GP ベースの DA の Stage2 から、判別得点値(d*)とパラメータ評価値 $\beta_i (i=1,\cdots,m)$ を得る。これらを用いて全観測値を以下に分類する。

$$G_0^T = \left\{ j \in G_0 \middle| \sum_{i=1}^{m} \beta_i^* x_{ij} \leq d^* \right\}, G_0^F = \left\{ j \in G_0 \middle| \sum_{i=1}^{m} \beta_i^* x_{ij} > d^* \right\}$$

$$G_1^T = \left\{ j \in G_1 \middle| \sum_{i=1}^{m} \beta_i^* x_{ij} > d^* \right\}, G_1^F = \left\{ j \in G_1 \middle| \sum_{i=1}^{m} \beta_i^* x_{ij} \leq d^* \right\}$$

"T"は、$G_h (h = 0 or 1)$ に正しく分類されたもの、"F"は間違って分類されたものを表している($\delta = 1$)。

Step2： $\sum_{i=1}^{m} \beta_i x_{ij}$ のä 番目のパラメータを $\beta_\delta = 0$ として GP ベースの GP を解き、判別得点値(d*)とパラメータ評価値 $\beta_i (i=1,\cdots,m)$ を得る。全観測値は次のように分けられる。

$$G_0^T(\delta) = \left\{ j \in G_0 \middle| \sum_{i \neq \delta}^{m} \beta_i^{**} x_{ij} \leq d^{**} \right\}, G_0^F = \left\{ j \in G_0 \middle| \sum_{i \neq \delta}^{m} \beta_i^{*} x_{ij} > d^{**} \right\}$$

$$G_1^T(\delta) = \left\{ j \in G_1 \middle| \sum_{i \neq \delta}^{m} \beta_i^{**} x_{ij} > d^{**} \right\}, G_1^F = \left\{ j \in G_1 \middle| \sum_{i \neq \delta}^{m} \beta_i^{**} x_{ij} \leq d^{**} \right\}$$

Step 3：統計値(Z)は以下のように計算され、自由度 1 で χ^2 分布に従う。

$$Z = \frac{\left\{ \#\left[G_0^T\right] - \#\left[G_0^T(\delta)\right] \right\}^2}{\#\left[G_0^T\right]} + \frac{\left\{ \#\left[G_1^T\right] - \#\left[G_1^T(\delta)\right] \right\}^2}{\#\left[G_1^T\right]} + \frac{\left\{ \#\left[G_0^F\right] - \#\left[G_0^F(\delta)\right] \right\}^2}{\#\left[G_0^F\right]} + \frac{\left\{ \#\left[G_1^F\right] - \#\left[G_1^F(\delta)\right] \right\}^2}{\#\left[G_1^F\right]}$$

Step 4：$H \geq \chi_\alpha^2$ ならば帰無仮説 $H_0 : \beta_\delta = 0$ は有意水準 α (片側)で棄却される。$\delta = m$ の場合は終了し、$\delta = \delta + 1$ なら Step2 へ戻る。

Step2 から Step4 までの計算過程は全てのパラメータ $\beta_i (i=1,\cdots,m)$ について行い、m回繰り返される。

5. 結論と今後の展望

本論文では、GP ベースの DA の新しい 2 つのノンパラメトリック検定を示し、GP ベースの DA を統計的に発展させた。1 つ目の検定は、観測値の 2 グループが同じ母集団を持つかどうかを統計的に調べ、2 つ目の検定は誤判別リスクを最小化するために各パラメータの重要度を調べるものである。また本論文では提案した GP ベースの DA と他の手法の優劣は比較せず、実用的・統計的な点を言及した。GP ベースの DA は、ロジット・プロフィットモデルに匹敵する数学的手法となるといえよう。

今後、提案した 2 つのノンパラメトリック検定は、実証研究へと発展させることができる。

<参考文献>

[1]Altman,E.I.(1968) "Financial Ratios, Discriminant Analysis and the Prediction Corporate Bankruptcy." *Journal of Finance*, 23, 589-609.

© 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.