統計数理研究所彙報
第32巻 第2号 (1984)

特別研究員研究報告

予測変量として連続変量と離散変量の混在を許す判別分析法

マッギル大学 高 根 芳 雄

これまで種々の判別分析の手法が開発されてきている (Lachenbruch, 1975; Hand, 1981). 予測変量に多次元正規性を仮定できる場合には正準変量に基づいた方法が広汎に用いられている (Fisher, 1936). これに対し予測変量が離散変量の場合はいまだ決定的な手法が存在しない (Goldstein and Dillon, 1978). 予測変量が離散変量の場合でも分布の仮定を離れ, 正準変量に基づいた方法が適用されることもあるが (Fisher, 1948; Hayashi, 1952) 記述統計学の域を出ない. 離散データの判別に対数線型モデルを使うことも考えられるが (Andersen, 1980; Sakamoto, 1982), このモデルでは逆に連続変量の取り扱いが困難である. 本報告では予測変量として連続変量と離散変量が混ざっている場合にも適用でき, なおかつ種々の統計的推測が可能であるような判別分析の手法を提案する.

この方法は次の様な仮定から成り立つ.

(a) 個々の観測個体は多次元ユークリッド空間の点として表わされる. 点の座標は予測変量の一次結合によって与えられる.

(b) 基準となる集団を代表する点 (その集団の最も典型的なものを表わす) も同じ空間の中に位置づけられる. その点はそれぞれの集団に属する個体の重心として与えられる.

(c) ある個体がある集団に所属する確率はその個体と集団間の距離の減少関数として規定される.

1. 方　　法

予測変量上の観測パターンを k で表わす. k の集団 α における観測頻度を $f_{k\alpha}$ とする. 予測変量上の観測値全体を G で表わす. 離散変量はすでにダミー変数化されているものとし, 連続変量は規準化されているものとする. Y を観測個体の座標を表わす行列, X を予測変量にかかる係数の行列とすると, (a) の仮定から

(1) $$Y = GX$$

が得られる. M を基準集団の重心の座標を表わす行列とすると

(2) $$M = (H'H)^{-1}H'Y = (H'H)^{-1}H'GX$$

と表わされる. ここで H は基準集団を表わすダミー変数の行列である. これより観測パターン k を持つ個体と集団 α 間のユークリッド距離の二乗は Y と M の適当な要素 $y_{k\alpha}$, $\mu_{\alpha\alpha}$ を用いて

(3) $$d^2_{k\alpha} = \sum_{a=1}^{A} (y_{k\alpha} - \mu_{\alpha\alpha})^2$$

と表わされる. A は予め指定された空間の次元数を表わす.

観測パターン k を持つ個体が集団 α に属する確率を $p_{k\alpha}$ とする. この $p_{k\alpha}$ が集団 α の先験

確率 p_α と，k と α の距離 $d_{k\alpha}$ を用いて

(4) $$p_{k\alpha}=\frac{p_\alpha \exp(-d_{k\alpha}^2)}{\sum_\beta p_\beta \exp(-d_{k\beta}^2)}$$

と表わされることを仮定する．この仮定は $p_{k\alpha}$ が $p_\alpha\cdot\exp(-d_{k\alpha}^2)$ に比例するという仮定から導かれる．即ち $c(\neq 0)$ について $p_{k\alpha}=cp_\alpha\exp(-d_{k\alpha}^2)$．ところが $\sum_\beta p_{k\beta}=1$ が成り立たなければならないから，結局 $c=(\sum_\beta p_\beta\exp(-d_{k\beta}^2))^{-1}$ が得られる．これは(4)の分母が単に規準化のファクターであることを示している．このモデルは $p_{k\alpha}$ が p_α に比例すること，$p_{k\alpha}$ が $d_{k\alpha}$ が増大するにつれ，$\exp(-d_{k\alpha}^2)$ に比例して減少する(距離の減少関数である)ことを表わしている．

これよりデータ全体の尤度は

(5) $$L=\prod_k\prod_\alpha(p_{k\alpha})^{f_{k\alpha}}$$

と表わされる．パラメータの推定値が最尤法によって求まったら，それを用いて $p_{k\alpha}$ を計算し，最大事後確率の原理に従って観測パターンを分類すればよい．

上記のモデルでは $\exp(-d_{k\alpha}^2)$ の部分はデータを構成する集団の大きさによって影響されない．従って，集団 α の周辺度数 f_α が予め指定されている場合(これを分離抽出という)も全体の標本の大きさ f だけが指定されている場合（同時抽出）と同様に取り扱うことができる．ただし，前者の場合は観測パターンを分類する段階で $p_{k\alpha}$ を計算するときに f_α/f のかわりに予め指定された何か別の p_α を用いなければならない．これは分離抽出の場合 f_α/f が母集団の大きさを反映しているとは限らないからである．

2. X にかかる制約

予測変量の種類の違いは係数にかかる制約条件の違いによって区別する．

(a) 変数が順序の付かないカテゴリー変数のとき：各カテゴリーについて A 次元の係数を求める．ただし

(6) $$\sum_j f_{i(j)}x_{i(j)a}=0$$

を満たさなければならない．$f_{i(j)}$ は変数 i のカテゴリー j の周辺頻度，$x_{i(j)a}$ は同じカテゴリーにかかる a 番目の係数である．(6)の制約は離散変量間の線型従属性に対処するためのものである．

(b) 変数が順序の付いたカテゴリー変数のとき：所与の順序条件を満たす一次元の係数（カテゴリーの数量化）を求め，それを多次元的に重み付ける．

(c) 連続変量のとき：すでに数量化が行われているものと仮定し，多次元の重み付けだけを求める．

(5)で定義された L をこれらの制約条件のもとで最大化する．

3. 考　　察

既に述べたように本文で提案された方法は連続変量も離散変量も同時に扱えるという利点がある．また標本が分離抽出の場合でも同時抽出と同様に扱える点も強味である．AIC (Akaike, 1974) を用いて最適な次元数を定めたり，予測変量を選択したりすることもできる．また連続変量をそのまま使った方がよいか，或いは離散化した方がよいかといった問題にも対処することができる．モデルは柔軟で，例えば距離関数を適当に変更することによってより一般的なモ

デルを作ることも可能である．

(4)のモデルは何人かの人によってほぼ同時に提案されたロジスティック判別分析（Cox, 1966; Anderson, 1972）に似ている．事実(4)のモデルで，常に基準集団数−1に相当する次元数をとるならばロジスティック判別分析に還元することが容易に証明できる．しかしながら常に次元数−1の次元数をとらなければならない必然性はなく，場合によってはかえってそれが害になることすらある．ところがロジスティック判別法では次元数（パラメータ数）を減らす可能性が全く考慮されていない．これに対し(4)のモデルではデータに則して最適な次元数が定められるという利点がある．

参 考 文 献

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on automatic control,* **19**, 716-723.

Andersen, E.B. (1980). *Discrete statistical models with social science applications,* North-Holland, Amsterdam.

Anderson, J.A. (1972). Separate sample logistic discrimination, *Biometrika,* **59**, 19-35.

Cox, D.R. (1966). Some procedure connected with the logistic qualitative response curve, (ed. David, F.N.), *Research papers in statistics Festschrift for J. Neyman*, Wiley, New York, 55-71.

Fisher, R.A. (1948). *Statistical methods for research workers,* Oliver and Boyd, (7th printing), London.

Fisher, R.A. (1936). The use of multiple measurement in taxonomic problems, *Annals of Eugenics,* **7**, 179-188.

Goldstein, M. and Dillon, W.R. (1978). *Discrete discriminant analysis,* Wiley, New York.

Hand, D.J. (1981). *Discrimination and classification*, Wiley, Chicheiter.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, *Annals of the Institute of Statistical Mathematics,* **2**, 69-98.

Lachenbruch, P.A. (1975). *Discriminant analysis*, Hafner, New York.

Sakamoto, Y. (1982). Efficient use of Akaike's information criterion for model selection in high dimensional contingency table analysis, *Metron,* **40**, 257-275.