論 文 審 査 の 要 旨

博士の専攻分野の名称	博　士（　理学　）	氏名	中　川　　智　之
学位授与の要件	学位規則第 4 条第 (1)・2 項該当		

論文題目

Estimating the probabilities of misclassification using CV when the dimension and the sample sizes are large　(高次元大標本の場合での CV を用いた誤判別確率の推定)

論文審査担当者

主　　査	教　授	若木　宏文
審査委員	教　授	井上　昭彦
審査委員	教　授	栁原　宏和

〔論文審査の要旨〕

判別問題とは, 複数の母集団の各々から得られた標本に関する観測値と, これらの母集団のいずれかに所属する個体に関する観測値が得られているとき, 個体がどの母集団に所属するかを判別 (判定) する問題である。本論文では, 2 つの母集団 Π_1, Π_2 に関する判別問題を扱っている。母集団 Π_k $(k = 1, 2)$ からの大きさ N_k の標本に関する観測値ベクトルを $\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, \ldots, \boldsymbol{x}_{k\,N_k}$ とする。これら, $N_1 + N_2$ 個の観測値ベクトルをまとめて訓練データと呼ぶ。判別対象である個体の判別は, 一般に, 訓練データから判別関数 d を作成し, 判別対象に関する観測値ベクトル $\boldsymbol{x}$ における $d(\boldsymbol{x})$ の値の符号によって以下のように行われる。

$$d(\boldsymbol{x}) > 0 \Rightarrow \text{ 母集団 } \Pi_1 \text{ に所属すると判定}$$
$$d(\boldsymbol{x}) \leq 0 \Rightarrow \text{ 母集団 } \Pi_2 \text{ に所属すると判定}$$

判別対象が Π_k $(k = 1, 2)$ に所属するときの分布を $\Pr(* \,|\, \boldsymbol{x} \in \Pi_k)$ と表すとき, 所属を誤って判別してしまう確率 $P(1|2) = \Pr(d(\boldsymbol{x}) > 0 \,|\, \boldsymbol{x} \in \Pi_1)$, $P(2|1) = \Pr(d(\boldsymbol{x}) \leq 0 \,|\, \boldsymbol{x} \in \Pi_2)$ を誤判別確率と呼ぶ。誤判別確率は判別方法 (判別関数) の良さの基準の一つであるが, Π_1, Π_2 の母集団分布に依存しており, これらは通常未知であるので推定の必要がある。

誤判別確率の推定方法として, 母集団分布にパラメトリックモデルを想定し, 判別関数の分布の漸近展開近似を利用する方法と, パラメトリックモデルを想定せずに, 訓練データの一部を検証用に取っておいて, 残りの訓練データから判別関数を作成し, 取っておいた検証用データが誤って判別される比率を用いる方法がある。クロスバリデーション (CV) とは, 訓練データから, i 番目の観測ベクトル $\boldsymbol{x}_{ik}$ を取り出し, 残りの $N_1 + N_2 - 1$ 個の訓練データから判別関数 ($d^{(-i)}$ とする) を作成し, $\boldsymbol{x}_{ik}$ が誤判別されるかどうかを調べる, という手順を $N_1 + N_2$ 回繰り返して, 誤判別の比率で推定する方法である。したがって, $P(2|1)$ の CV 推定値は,

$$\hat{P}_{\mathrm{CV}}(2|1) = \frac{1}{N_1} \sum_{i=1}^{N_1} 1(d^{(-i)}(\boldsymbol{x}_{i1}) \leq 1)$$

となる。ただし, 1(不等式) は不等式が真のとき 1, 偽のとき 0 である。

観測ベクトルの次元を p とする。フィッシャーの線形判別関数について, p を固定し, $N_1, N_2 \to \infty$ (大標本漸近枠組み) の下での漸近展開近似に基づく誤判別確率の推定量が McLachlan (1974) によって提案されているが, p が大きくなると推定精度が悪くなることが知られている。 参考文献 (1) では, N_1, N_2 に加えて p も大きくなる場合 (高次元・大標本漸近枠組み) の下での漸近展開近似を利用した誤判別確率の推定量が提案されている。参考論文 (2) では, 大標本漸近枠組みの下での漸近展開近似を利用した, 線形判別関数と 2 次判別関数の選択方法を提案している。

クロスバリデーション推定量について, 大標本漸近枠組みでは N_1^{-1}, N_2^{-1} に関して 2 次の漸近不偏性が知られているが, 高次元・大標本漸近枠組みによる漸近不偏性に関する理論研究はなされていなかった。

本論文では, 高次元・大標本漸近枠組みの下では, クロスバリデーション推定量は, $N_1^{-1}, N_2^{-1}, p^{-1}$ に関して 1 次までしか漸近不偏性がないことを示し, 2 次までの漸近不偏性を得るため方法として

(1) 検証用に 2 個取っておくクロスバリデーション推定量を利用するバイアス補正法

(2) 検証用の観測ベクトルの情報も判別関数の構成に少しだけ利用するバイアス補正法

(3) 判別関数の閾値をずらすことによるバイアス補正法

を提案し, 高次元・大標本漸近展開近似に基づく誤判別確率の推定量, 従来のクロスバリデーションとの優劣を数値実験を通して明らかにした. バイアス補正としては, 手法 (1), (2), (3) は, ほぼ同等であった。 手法 (1) は分散が大きくなるため平均 2 乗誤差が大きくなってしまうが, 手法 (2), (3) と漸近展開を用いた手法は分散が大きくならないため, 平均 2 乗誤差の意味でも有効な手法といえる。

以上, 審査の結果, 本論文の著者は博士（理学）の学位を授与される十分な資格があるものと認める。

参考論文

(1) EPMC estimation in discriminant analysis when the dimension and sample sizes are large.

T. Tonda, T. Nakagawa and H. Wakaki, *Hiroshima Mathematical Journal*, **47**(1), (2017), 43-62.

(2) Selection of the linear and the quadratic discriminant functions when the differece beween two covariance matrices is small.

T. Nakagawa and H. Wakaki, *Journal of the Japan Statistical Society*, **47**(2), (2017), 145-165.