多変量関数部分空間法に基づく高次元経時測定データの分類とその応用

Multivariate functional subspace classification for high-dimensional longitudinal data and its application

数学専攻　福田 竜也

FUKUDA, Tatsuya

1　はじめに

近年, 高次元経時測定データの分類問題は, 医学研究や気象学, 生態学など諸科学の様々な分野で研究が行われている. 経時測定データは, 平滑化処理を施すことで関数データとして扱うことができる. 本論文では, 関数データに対する判別器の構築を行い, 経時測定データに対する適用を議論する. 経時測定データの平滑化は, 現象の真の構造を表す関数が基底関数の線形結合で表されていると仮定する基底展開法を用いて行われ, 基底関数としては B-スプラインや Bernstein 基底関数などが使われる. 個体ごとに観測時点やその総数が異なるような経時測定データには多変量データに対する解析手法の適用が難しいことから, 関数データ解析を用いたアプローチが研究されている.

また, 多変量データの分類手法として部分空間法がある. 部分空間法は, 多変量データから構成される低次元の部分空間に着目して判別を行う手法で, 主成分分析によって次元圧縮を行う CLAFIC (CLAss-Featuring Information Compression) 法 (Watanabe, 1969) が代表的な手法として知られている. CLAFIC 法はパターン変動を捉え, 少ない学習データ数でも安定し, かつ低い計算コストで判別を行う分類手法として知られ, 音声認識や画像判別, 文字判別などで用いられている.

多変量データに対する分類手法を拡張することで, 関数データの枠組みで分類問題を考えることができる. 関数データの分類手法として, 例えば, 関数線形判別分析 (James and Hastie, 2001) や関数サポートベクターマシン (Rossi and Villa, 2006) がある. 多群かつ多次元の関数データの分類を行う際には, 計算コストの低さや少ない学習データ数でも判別することができる判別器の構成が求められる. 本論文では, これらの点に有効に働く CLAFIC 法を拡張し, 関数主成分分析 (Besse and Ramsay, 1986) による次元圧縮を用いて関数データの分類を行う多変量関数部分空間法を提案する. 多変量関数部分空間法は, 関数主成分分析による次元圧縮によって構成された部分空間と, クラスが未知の観測データとの類似度を測ることで判別を行う.

2　関数主成分分析

関数主成分分析 (Functional principal component analysis; FPCA) は, Besse and Ramsay (1986) により提唱され, 基底展開法に基づいて平滑化された関数データの集合に対して, 主成分分析を施す手法である.

いま, p 個の特性に関して観測された n 個体の経時測定データを, 基底展開法に基づく非線形回帰モデリングにより平滑化した p 次元関数データ集合を $\{x_{i1}(t), x_{i2}(t), \cdots, x_{ip}(t)\ ;\ t \in \mathcal{T}\}\ (i = 1, 2, \cdots, n)$ とする. ただし, i 番目の個体の l 番目の特性に関する関数データは次のように表されているとする.

$$x_{il}(t) = \sum_{m=1}^{M_l} \hat{\gamma}_{ilm} b_{lm}(t) = \hat{\boldsymbol{\gamma}}_{il}' \boldsymbol{b}_l(t)\ ;\ t \in \mathcal{T}, \quad l = 1, 2, \cdots, p, \quad i = 1, 2, \cdots, n.$$

ここで, $\hat{\boldsymbol{\gamma}}_{il} = (\hat{\gamma}_{il1}, \hat{\gamma}_{il2}, \cdots, \hat{\gamma}_{ilM_l})'$ は推定された M_l 次元回帰係数ベクトル, $\boldsymbol{b}_l(t) = (b_{l1}(t), b_{l2}(t), \cdots, b_{lM_l}(t))'$

は基底関数ベクトルである．また, 重み関数ベクトル $\boldsymbol{w}(t) = (w_1(t), w_2(t), \cdots, w_p(t))'$ の l 番目の特性に関する重み関数 $w_l(t)$ は次のように, 関数データ $x_{il}(t)$ と同じ基底関数の線形結合で与えられているとする．

$$w_l(t) = \sum_{m=1}^{M_l} \beta_{lm} b_{lm}(t) = \boldsymbol{\beta}_l' \boldsymbol{b}_l(t), \qquad l = 1, 2, \cdots, p.$$

ただし, $\boldsymbol{\beta}_l = (\beta_{l1}, \beta_{l2}, \cdots, \beta_{lM_l})'$ である．このとき, i 番目の個体の関数データベクトル $\boldsymbol{x}_i(t)$ と重み関数ベクトル $\boldsymbol{w}(t)$ の内積 f_i を次の式で定義する．

$$f_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle = \sum_{l=1}^{p} \int_{\mathcal{T}} w_l(t) x_{il}(t) dt = \sum_{l=1}^{p} \int_{\mathcal{T}} \boldsymbol{\beta}_l' \boldsymbol{b}_l(t) \boldsymbol{b}_l(t)' \hat{\boldsymbol{\gamma}}_{il} dt = \sum_{l=1}^{p} \boldsymbol{\beta}_l' W^l \hat{\boldsymbol{\gamma}}_{il} = \boldsymbol{\beta}' W \hat{\boldsymbol{\gamma}}_i. \tag{1}$$

ただし, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \cdots, \boldsymbol{\beta}_p')'$, $\hat{\boldsymbol{\gamma}}_i = (\hat{\boldsymbol{\gamma}}_{i1}', \hat{\boldsymbol{\gamma}}_{i2}', \cdots, \hat{\boldsymbol{\gamma}}_{ip}')'$ であり, $W^l = \int_{\mathcal{T}} \boldsymbol{b}_l(t) \boldsymbol{b}_l(t)' dt$ は, その (r, s) 要素に

$$W_{rs}^l = \int_{\mathcal{T}} b_{lr}(t) b_{ls}(t) dt, \qquad r, s = 1, 2, \cdots, M_l$$

を持つ $M_l \times M_l$ 対称行列で, ここでは交差積行列と呼ぶ．ここで, $M = M_1 + M_2 + \cdots + M_p$ とすると, $M \times M$ 行列 W は, 交差積行列を対角成分に並べた行列 $W = \mathrm{diag}(W^1, W^2, \cdots, W^p)$ として定義され, これを p 次交差積行列と呼ぶ．このとき, (1) 式より, $f_1, f_2, \cdots, f_n$ の標本分散 s_f^2 は次のようになる．

$$s_f^2 = \frac{1}{n} \sum_{i=1}^{n} (f_i - \overline{f})^2 = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}' W (\hat{\boldsymbol{\gamma}}_i - \overline{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}}_i - \overline{\boldsymbol{\gamma}})' W \boldsymbol{\beta} = \boldsymbol{\beta}' W \Gamma W \boldsymbol{\beta}. \tag{2}$$

ただし, $\overline{\boldsymbol{\gamma}} = n^{-1} \sum_{j=1}^{n} \hat{\boldsymbol{\gamma}}_j$ を推定係数ベクトルの標本平均ベクトルとし, 関数データベクトル $\boldsymbol{x}_i(t)$ の中心化された推定係数ベクトルの $M \times M$ 標本分散共分散行列を $\Gamma = n^{-1} \sum_{i=1}^{n} (\hat{\boldsymbol{\gamma}}_i - \overline{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}}_i - \overline{\boldsymbol{\gamma}})'$ とした．(2) 式の標本分散を, 制約条件 $\langle \boldsymbol{w}_j, \boldsymbol{w}_k \rangle = \boldsymbol{\beta}_j' W \boldsymbol{\beta}_k = \delta_{jk}$ のもとで最大化する．主成分分析と同様に, ラグランジュの未定乗数法により分散の最大化問題は行列の固有値問題に帰着される．そこで, 行列 $W\Gamma W$ の固有値問題を解いて得られる M 個の固有値を $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_M$, とし, 各固有値に対応する固有ベクトルを $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}', \boldsymbol{\beta}_{k2}', \cdots, \boldsymbol{\beta}_{kl}', \cdots, \boldsymbol{\beta}_{kp}')'$ $(k = 1, 2, \cdots, M)$ とすると, 第 k 主成分関数ベクトル $\boldsymbol{w}_k(t)$ と, l 番目の特性に関する主成分関数 $w_{kl}(t)$ は次で与えられる．

$$\begin{aligned} \boldsymbol{w}_k(t) &= (w_{k1}(t), w_{k2}(t), \cdots, w_{kp}(t))', \qquad k = 1, 2, \cdots, M, \\ w_{kl}(t) &= \sum_{m=1}^{M_l} \beta_{klm} b_{lm}(t) = \boldsymbol{\beta}_{kl}' \boldsymbol{b}_l(t). \end{aligned}$$

3 多変量関数部分空間法

3.1 関数データに対する CLAFIC 法

いま, K 個のクラス $G_1, G_2, \cdots, G_K$ があり, k 番目のクラス G_k の p 次元関数データ集合を $\{\boldsymbol{x}_1^{(k)}(t), \boldsymbol{x}_2^{(k)}(t), \cdots, \boldsymbol{x}_{n_k}^{(k)}(t)\ ;\ t \in \mathcal{T}\}$ で与える．この関数データ集合に対し関数主成分分析を施し, 主成分関数ベクトル $\boldsymbol{w}_1^{(k)}(t), \boldsymbol{w}_2^{(k)}(t), \cdots, \boldsymbol{w}_M^{(k)}(t)$ を得る．多変量関数部分空間法では, $d_k (< M)$ 個の主成分関数ベクトル $Z_k^* = \left(\boldsymbol{w}_1^{(k)}(t), \boldsymbol{w}_2^{(k)}(t), \cdots, \boldsymbol{w}_{d_k}^{(k)}(t)\right)$ によって構成される部分空間を考える．各クラスに対し構成した部分空間を $\mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_K$ とする．通常の CLAFIC 法では, 各部分空間に対してクラスが未知のデータを射影行列によって射影し, 正射影の長さを類似度としてデータの判別を行っていた．しかし, 関数データの場合,

これらの部分空間およびクラスが未知のデータは時間 $t \in \mathcal{T}$ に依存する多様体であるから, 部分空間への射影について議論する必要がある. 時間に依存するデータを扱う場合, 次のように考えればよい.

まず, ある時間 $t_a \in \mathcal{T}$ に固定して考えてみると, $Z_k^*|_{t=t_a}$ は実数値ベクトルからなる行列になり, $\mathcal{L}_k|_{t=t_a}$ への正射影とその長さを求めることが可能になる. すなわち, 時間を固定することで通常の部分空間法の適用が可能になる. そこで, ある時間に固定し観測データの正射影およびその長さを求めるプロセスを, すべての $t \in \mathcal{T}$ ついて行い, 正射影の長さを足し合わせた値を用いて, 判別を行う. クラスが未知の観測 p 次元関数データを $\boldsymbol{z}(t)$ とすると, ある時間 $t \in \mathcal{T}$ における部分空間 $\mathcal{L}_k$ への正射影は Karhunen-Loéve 展開を用いて

$$\boldsymbol{z}^{(k)}(t) = \sum_{j=1}^{d_k} \langle \boldsymbol{z}, \boldsymbol{w}_j^{(k)} \rangle \boldsymbol{w}_j^{(k)}(t) \tag{3}$$

と表せる. この正射影の長さ, すなわち (3) 式で与えられるベクトルの各成分の二乗和をすべての $t \in \mathcal{T}$ について足し合わせたものを類似度として用いる. ゆえに観測データのクラスを判別する類似度を次で定義する.

$$S_k(\boldsymbol{z}(t)) \equiv \int_{\mathcal{T}} \sum_{l=1}^{p} \left(z_l^{(k)}(t) \right)^2 dt = \sum_{l=1}^{p} \int_{\mathcal{T}} \left(z_l^{(k)}(t) \right)^2 dt = \|\boldsymbol{z}^{(k)}\|^2. \tag{4}$$

ただし, $\boldsymbol{z}^{(k)}(t)$ $(t \in \mathcal{T})$ は部分空間 $\mathcal{L}_k$ への正射影であり, $\boldsymbol{z}^{(k)}(t) = \left(z_1^{(k)}(t), z_2^{(k)}(t), \cdots, z_p^{(k)}(t) \right)'$ とする. さらに, (4) 式で定義した類似度は, 重み関数ベクトル $\boldsymbol{w}_j^{(k)}(t)$ の性質を用いて次のように書き換えることができる. $f_j^k = \langle \boldsymbol{z}, \boldsymbol{w}_j^{(k)} \rangle$ とおくと

$$\begin{aligned} S_k(\boldsymbol{z}(t)) = \|\boldsymbol{z}^{(k)}\|^2 &= \sum_{l=1}^{p} \int_{\mathcal{T}} \left(\sum_{j=1}^{d_k} f_j^k w_{jl}^{(k)}(t) \right) \left(\sum_{h=1}^{d_k} f_h^k w_{hl}^{(k)}(t) \right) dt \\ &= \sum_{l=1}^{p} \int_{\mathcal{T}} \left(\sum_{j=1}^{d_k} (f_j^k)^2 \left(w_{jl}^{(k)}(t) \right)^2 + 2 \sum_{j<h} f_j^k f_h^k \left(w_{jl}^{(k)}(t) w_{hl}^{(k)}(t) \right) \right) dt \\ &= \sum_{j=1}^{d_k} (f_j^k)^2. \end{aligned} \tag{5}$$

ただし, 重み関数ベクトルが満たす次の条件を用いた.

$$\|\boldsymbol{w}_j^{(k)}\|^2 = \sum_{l=1}^{p} \int_{\mathcal{T}} \left(w_{jl}^{(k)}(t) \right)^2 dt = 1, \quad \langle \boldsymbol{w}_j^{(k)}, \boldsymbol{w}_h^{(k)} \rangle = \sum_{l=1}^{p} \int_{\mathcal{T}} w_{jl}^{(k)}(t) w_{hl}^{(k)}(t) dt = 0 \ \ (j \neq h).$$

(5) 式により, 類似度を主成分スコアの二乗和で求めることができる. この計算方法では重み関数自体を求める必要がなく, 重みおよび観測された関数データの係数ベクトルと交差積行列のみで類似度が求まるため, 計算コストの低下が期待できる. この類似度を用いて, 次の判別規則が定まる.

$$r = \underset{k=1,2,\cdots,K}{\arg\max}\ S_k(\boldsymbol{z}(t)) \implies \boldsymbol{z}(t) \in G_r.$$

すなわち, クラス毎に類似度を計算し, 類似度が最も大きいクラスへ観測データを判別する.

3.2 部分空間の次元の選択

部分空間 $\mathcal{L}_k$ の次元数 $d_k (< M)$ の選択は非常に重要な問題である. 関数部分空間法における判別率は, 部分空間の次元数に強く依存しており, 一般に, 部分空間の次元を高くしすぎると, 部分空間同士の重なりが増加す

ることから, 判別性能が低下してしまう. 逆に次元が低すぎると, 各クラスの近似精度が低下するため結果として判別性能は低下する. そこで, 代表的な例として累積寄与率を用いる方法が挙げられる. これは各部分空間ごとに累積寄与率を計算し, その値がある閾値 α を初めて超えたときの次元数を選択する方法である. ここで, クラス G_k に対する第 d_k 主成分までの累積寄与率は次で与えられる.

$$a(d_k) = \frac{\rho_{k1} + \rho_{k2} + \cdots + \rho_{kd_k}}{\rho_{k1} + \rho_{k2} + \cdots + \rho_{kd_k} + \cdots + \rho_{kM}}.$$

ただし, $\rho_{k1}, \rho_{k2}, \cdots, \rho_{kM}$ はクラス G_k に対して計算された行列 $W\Gamma_k W$ の固有値である. このとき $a(d_k - 1) < \alpha \leq a(d_k)$ を満たす d_k を, クラス G_k の部分空間の次元として選択する. しかし, 閾値の設定は解析者の主観や経験に基づく部分があり, 客観的な選択であるとはいえない. さらに, 主成分分析にはクラスという概念がないため, 閾値をうまく設定することで判別率が上がるといった保証はない. そこで, 本論文では層化 v-分割クロスバリデーションによって部分空間の次元数を決定する.

層化 v-分割クロスバリデーションでは, 各クラスの学習データを v 分割し, 得られた互いに素な各クラス v 個のデータ集合に対し, 1 つを判別モデルの評価用データ, それ以外を判別モデルの学習データとして判別モデルの精度を測る. このプロセスを v 個すべての集合について行い, 平均的な評価値 CV_{error} を計算する. この CV_{error} を最小にするような次元数 d_k を, 最適な次元数として選択する. 詳細は, Holmström *et al.* (1996) を参照されたい.

4 おわりに

本論文では, 経時測定データの分類を行う多変量関数部分空間法を提案し, 数値実験と手書き数字のデータの判別を通して提案手法の有用性を検証した. 今後の展望としては, 学習データ数が少なく観測時点数やその総数が個体ごとに異なるようなデータへ提案手法を適用し, 有効性を検証することが挙げられる. また, 主成分分析は教師あり学習に特化した次元圧縮法であるとはいえない. そこで, 各クラスが固有に持つ特徴量を抽出するような次元圧縮手法を用いて関数部分空間法に適用していくことが今後の研究課題である.

参考文献

[1] Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* 51(2), 285-311.

[2] Holmström, L., Koistinen, P., Laaksonen, J. and Oja, E. (1996). Comparison of Neural and Statistical Classifiers - Theory and Practice. Research Reports A13, Rolf Nevanlinna Institute, University of Helsinki, Finland.

[3] James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society* Series B, 63:533-550

[4] Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7-9), 730-742.

[5] Watanabe, S. (1969). *Knowing and Guessing Quantitative Study of Inference and Information*. John Wiley and Sons.