5F-05

観測ベクトルの成分の重要度を推定して 2 群間の差異を増幅する次元圧縮法とその生体信号解析への応用 *

小谷野 仁 †
農業・食品産業技術総合研究機構

1. はじめに

本稿では, 生体信号に基づいた個体の分類問題を考える. 考察する問題は, 詳しくは第 3 節で述べるが, 形式的には次の問題として定式化することができる. (i) 2 つの群のうちのどちらに属するかが既知であるいくつかの個体から, 全ての個体において同一のいくつかの時点 (例えば年齢) において観測ベクトルが収集されている. (ii) 但し, 観測ベクトルの成分には, 個体の帰属と無関係なものが含まれている. (iii) そのような成分は全ての個体と時点において共通であるが, (iv) それらがどれかは未知である. この時, 帰属が未知である個体から, 教師データの観測の時点と同一のいくつかの時点において収集された観測ベクトルに基づいて, その個体を 2 つの群のうちの一方に分類せよ. 上記の (i) から (iv) は, 個体を分類するのに, 関係しているかどうかは分からないけれども, ひとまず持っている情報は全て持って来たという設定であり, データ解析においてよく起こる状況である.

分類問題ではなく, 連続変数の観測値を説明, 推定, または予測しようとする回帰の問題に対しては, 上記の (ii) から (iv) の設定の下でも様々な方法が存在する. 例えば, 線形回帰モデルの偏回帰係数を推定し, 係数が 0 であるという仮説の t 検定を行って, その仮説を棄却できない変数を説明変数から除いたり, Ridge 回帰や Lasso 回帰を用いることが考えられる.

また, (ii) から (iv) の設定のない分類問題に対しても, 判別分析, サポートベクターマシーン (以下では SVM と略記), SVM とカーネル法の組合せなどの色々な方法がある. 例えば, Fisher の判別分析は, 群間分散の群内分散に対する比を最小化する, より低い次元の部分空間に観測ベクトルを射影することにより, 2 つの群の母集団分布の重なりをできるだけ小さくして, 個体を高い精度で分類しようとするアプローチであった. 一方で, SVM とカーネル法を組み合わせる方法は, 観測ベクトルをより高い次元の空間に写像して, その空間に分離超平面を定めることにより, 元の空間に分離超平面を定めたのでは個体を高い精度で分類することが難しい場合でも, 高い精度の分類を実現しようとするアプローチであった.

SVM は伝統的にカーネル法と用いられてきた. この影響で, 多変量統計解析にもカーネル法を取り入れようとする研究がこれまでに色々となされてきた. 本稿では, このような研究の方向性とは逆に, 多変量解析における判別分析や主成分分析の基礎的な考え方である次元圧縮法を SVM と組み合わせる. 本稿で提案する分類方式は, これらに更にスペクトル解析と集団学習も組み合わせたごった煮の方法であるが, これらのうち新規に設計した次元圧縮法がキーになっている.

2. 提案する分類方式

まず, 本稿で提案する次元圧縮法を述べる. 教師データとして, 時点 $i = 1, \cdots, \ell$ において, 群 1 と 2 に属する m 個と n 個の個体からそれぞれ k 次元観測ベクトル

$$\boldsymbol{x}_{1,i} = (x_{1,i}^{(1)}, \cdots, x_{1,i}^{(k)}), \cdots, \boldsymbol{x}_{m,i} = (x_{m,i}^{(1)}, \cdots, x_{m,i}^{(k)}),$$
$$\boldsymbol{y}_{1,i} = (y_{1,i}^{(1)}, \cdots, y_{1,i}^{(k)}), \cdots, \boldsymbol{y}_{n,i} = (y_{n,i}^{(1)}, \cdots, y_{n,i}^{(k)})$$

が得られているとする. $\bar{\boldsymbol{x}}_i = (1/m)\sum_{h=1}^{m} \boldsymbol{x}_{h,i}$ と $\bar{\boldsymbol{y}}_i = (1/n)\sum_{h=1}^{n} \boldsymbol{y}_{h,i}$ とおき, 各 $j = 1, \cdots, k$ に対して, $\bar{\boldsymbol{x}}_i$ と $\bar{\boldsymbol{y}}_i$ の第 j 成分をそれぞれ $\bar{x}_i^{(j)}$ と $\bar{y}_i^{(j)}$ によって表す.

任意の $c > 0$ と各 $i = 1, \cdots, \ell$ に対して, $\bar{x}_i^{(j)} - \bar{y}_i^{(j)} > c$ となる上付き添え字の数を $r(i,c)$ によって表し, 各 $j = 1, \cdots, r(i,c)$ に対して, j 番目に大きいそのような上付き添え字を $\hat{j}(i,c,j)$ によって表す. また, $\bar{y}_i^{(j)} - \bar{x}_i^{(j)} > c$ となる上付き添え字の数を $s(i,c)$ によって表し, 各 $j = 1, \cdots, s(i,c)$ に対して, j 番目に大きいそのような上付き添え字を $\tilde{j}(i,c,j)$ によって表す. 以下では, $r(i,c), s(i,c), \hat{j}(i,c,j)$, 及び $\tilde{j}(i,c,j)$ をそれぞれ $r, s, \hat{j}(j)$, 及び $\tilde{j}(j)$ と略記する. 次

*Dimension reduction method that amplifies the difference between two groups by estimating the importance of the components of observation vectors and its application to biological signal analysis

†Hitoshi Koyano, National Agriculture and Food Research Organization

の最大化問題の解をそれぞれ $\alpha^*_{i,c,1},\cdots,\alpha^*_{i,c,r}$ と $\beta^*_{i,c,1},\cdots,\beta^*_{i,c,s}$ によって表す.

$$\max_{\alpha_1,\cdots,\alpha_r}\left(\sum_{h=1}^{m}\sum_{j=1}^{r}\alpha_j x_{h,i}^{(\hat{j}(j))}-\sum_{h=1}^{n}\sum_{j=1}^{r}\alpha_j y_{h,i}^{(\hat{j}(j))}\right),$$

$$\max_{\beta_1,\cdots,\beta_s}\left(\sum_{h=1}^{n}\sum_{j=1}^{s}\beta_j y_{h,i}^{(\tilde{j}(j))}-\sum_{h=1}^{m}\sum_{j=1}^{s}\beta_j x_{h,i}^{(\tilde{j}(j))}\right).$$

そうして,

$$\xi^{(1)}_{h,i,c}=\sum_{j=1}^{r}\alpha^*_{i,c,j}x_{h,i}^{(\hat{j}(j))},\ \xi^{(2)}_{h,i,c}=\sum_{j=1}^{s}\beta^*_{i,c,j}x_{h,i}^{(\tilde{j}(j))},$$

$$\eta^{(1)}_{h,i,c}=\sum_{j=1}^{r}\alpha^*_{i,c,j}y_{h,i}^{(\hat{j}(j))},\ \eta^{(2)}_{h,i,c}=\sum_{j=1}^{s}\beta^*_{i,c,j}y_{h,i}^{(\tilde{j}(j))}$$

と定め, $\boldsymbol{x}_{h,i}$ と $\boldsymbol{y}_{h,i}$ をそれぞれ

$$\boldsymbol{\xi}_{h,i,c}=(\xi^{(1)}_{h,i,c},\xi^{(2)}_{h,i,c}),\ \boldsymbol{\eta}_{h,i,c}=(\eta^{(1)}_{h,i,c},\eta^{(2)}_{h,i,c})$$

と次元圧縮する.

次に, 本稿で提案する分類方式を述べる. まず, 学習方式は次の通りである.

ステップ 1. $c>0$ と正の整数 f を定める.

ステップ 2. 各 $i=1,\cdots,\ell$ に対して, 上記の次元圧縮法で教師データ $\boldsymbol{x}_{1,i},\cdots,\boldsymbol{x}_{m,i}$ と $\boldsymbol{y}_{1,i},\cdots,\boldsymbol{y}_{n,i}$ から 2 群からの観測ベクトルの差異を増幅させる係数 $\alpha^*_{i,c,1},\cdots,\alpha^*_{i,c,r}$ と $\beta^*_{i,c,1},\cdots,\beta^*_{i,c,s}$ を求め, 特徴ベクトル $\boldsymbol{\xi}_{1,i,c},\cdots,\boldsymbol{\xi}_{m,i,c}$ と $\boldsymbol{\eta}_{1,i,c},\cdots,\boldsymbol{\eta}_{n,i,c}$ を作成する.

ステップ 3. 各 $i=1,\cdots,\ell$ に対して, $\boldsymbol{\xi}_{1,i,c},\cdots,\boldsymbol{\xi}_{m,i,c}$ と $\boldsymbol{\eta}_{1,i,c},\cdots,\boldsymbol{\eta}_{n,i,c}$ を SVM に与えて, ℓ 個の学習済み SVM $M_1,\cdots,M_\ell$ を作成する.

ステップ 4. 各 $i=1,\cdots,\ell$ に対して, 交差検証を行って M_i の正解率 a_i を計算する.

なお, 教師データが生体信号である場合, 予め Hamming 窓などの窓関数を掛け, Fourier 変換を施して, Fourier 変換の実部, 虚部, 及び絶対値の f 以下の周波数に関する和として, 各 $i=1,\cdots,3g$ に対する $\boldsymbol{x}_{1,i},\cdots,\boldsymbol{x}_{m,i}$ と $\boldsymbol{y}_{1,i},\cdots,\boldsymbol{y}_{n,i}$ を作成しておく. ここで, g は各個体から 1 回の測定時に収集する生体信号の数である.

次に, 分類方式は次の通りである. 群 1 と 2 のうちのどちらに属するかが未知である個体からの時点 $i=1,\cdots,\ell$ における観測ベクトル $\boldsymbol{z}_i=(z_i^{(1)},\cdots,z_i^{(k)})$ が与えられたとする.

ステップ 1. 各 $i=1,\cdots,\ell$ に対して, 学習方式のステップ 2 で求めた $\alpha^*_{i,c,1},\cdots,\alpha^*_{i,c,r}$ と $\beta^*_{i,c,1},\cdots,\beta^*_{i,c,s}$ を用いて

$$\zeta^{(1)}_{i,c}=\sum_{j=1}^{r}\alpha^*_{i,c,j}z_i^{(\hat{j}(j))},\ \zeta^{(2)}_{i,c}=\sum_{j=1}^{s}\beta^*_{i,c,j}z_i^{(\tilde{j}(j))}$$

を計算し, $\boldsymbol{\zeta}_{i,c}=(\zeta^{(1)}_{i,c},\zeta^{(2)}_{i,c})$ とおく.

ステップ 2. 各 $i=1,\cdots,\ell$ に対して, $\boldsymbol{\zeta}_{i,c}$ を M_i に与えて, 新規の個体を分類させる. $M_1,\cdots,M_\ell$ のうち新規の個体を群 1 と 2 に分類したものの添え字の集合をそれぞれ I_1 と I_2 によって表す.

ステップ 3. 不等式 $\sum_{i\in I_1}a_i>\sum_{i\in I_2}a_i$ が成り立つならば, 新規の個体を群 1 に分類し, そうでないならば, 群 2 に分類する.

3. 生体信号解析への応用

第 2 節で述べた分類方式を用いて, 集団中のマウスからいくつかの月齢においていくつかの音圧の下で聴性脳幹反応を測定し, 各マウスが将来のある時点において聴力がほぼ一定である群と大きく低下する群のうちのどちらに属するかを予測する問題に取り組んだ. 統計学や機械学習の通常の分類・予測問題と比較した時, この問題には次の特徴がある. 聴性脳幹反応に基づいて, それらを測定した時点のマウスの集団を 2 つの群に分けるのではなく, 将来の時点のマウスの集団を 2 つの群に分ける. 従って, この問題は分類問題であると同時に, 予測問題である.

[1] で述べたように, 第 2 節で提案した分類方式を少し簡便化したものによって, 87 個体のマウスの 3 カ月齢と 6 カ月齢における聴性脳幹反応に基づいて, それらのマウスが 9 カ月齢において聴力がほぼ一定であるか大きく低下するかを予測させ, 3 分割交差検証を 10000 回行ったところ, 0.9003 の正解率を得た. マウスの 3 カ月は人間の約 8 年に相当する. 3 カ月齢または 6 カ月齢の時点でその後の 6 カ月間または 3 カ月間の聴力の変動が完全に定まっているとは思えない. しかし, 9 カ月齢まで聴力がほぼ一定である集団と 9 カ月齢までに聴力が大幅に低下する集団の間には, 3 カ月齢と 6 カ月齢における聴性脳幹反応に僅かな差異があり, 第 2 節で提案した次元圧縮法はそれらを抽出, 増幅することができ, その結果, 提案した分類方式により聴力の変動を高い精度で予測できた可能性がある.

4. 本発表の目的

本発表では, 第 2 節で述べた次元圧縮法に与えた数理的基礎付けを述べる. また, 新規に作成した聴性脳幹反応のいくつかの模擬データセットを解析して, 第 2 節で述べた分類方式の実際のデータ解析における頑健性を調べた結果を紹介する.

引用文献

[1] 大池秀明, 小谷野仁. 聴性脳幹反応に基づいて将来の進行性難聴を予測する方式を構築する方法. 特願 2022-192092. 2022-11-30.