

松澤 知己† レラトー レイサ‡ 伊藤 栄祐† 加藤 毅†

† 群馬大学大学院理工学専攻 ‡ 産総研人工知能研究センター

1 はじめに

ブレインコンピュータインターフェイス (BCI) は、脳波信号を解析することによって、脳とコンピュータの直接的なコミュニケーションを可能にするための技術である。近年、脳波信号を表現するために、共分散記述子を用いる手法が注目されている [1]。共分散記述子は正定値行列であり、共分散記述子を使った解析には、アフィン不変距離、対数ユークリッド距離など、これまで様々な距離関数が試されてきた。一方、通常のベクトル空間を使ったパターン認識では、データから距離計量を学習することにより汎化性能が向上することが確かめられている [2]。本研究では、共分散記述子によるパターン認識の性能向上をはかるため、訓練用データから正定値錐上の距離を学習する方法を開発した。

正定値錐上の計量学習は本研究が初ではなく、ITML [2] を拡張した方法が 2015 年にすでに提案されていた [3]。しかし、その学習アルゴリズムは致命的な誤りを含んでいた。本稿では、その誤りを正すとともに、3 節において損失関数のさらなる一般化を行う。4 節では、最適化アルゴリズムにおいて、筆者らが新たに見つけた 2 つの発見を述べる：(i) その各反復を $O(n^3 + Ln)$ (n および L の定義は後述) の計算コストに抑えられる；(ii) 乱択戦略の導入により、反復回数を劇的に減少できる。5 節では、脳波信号の分類実験において、計量学習により汎化性能が顕著に向上することを実験的に示す。記法は、文献 [4] 参照。

2 正定値錐上の距離計量関数

従来、2 つの半正定値行列 $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{S}_+^n$ の距離計量関数として、 $D_{\Phi}(\mathbf{X}_1, \mathbf{X}_2) := \|\Phi(\mathbf{X}_1) - \Phi(\mathbf{X}_2)\|_F^2$ であらわされる距離関数が検討されてきた。写像 $\Phi: \mathbb{S}_+^n \rightarrow \mathbb{R}^{n \times m}$ には、 $\Phi(\mathbf{X}) := \text{logm}(\mathbf{X})$, $\Phi(\mathbf{X}) := \mathbf{X}$, $\Phi(\mathbf{X}) := \text{chol}(\mathbf{X})$ などが選ばれてきた。本研究では、次のパラメトリックな距離計量を導入する：

$$D_{\Phi}(\mathbf{X}_1, \mathbf{X}_2; \mathbf{W}) := \langle \mathbf{W}, (\Phi(\mathbf{X}_1) - \Phi(\mathbf{X}_2))(\Phi(\mathbf{X}_1) - \Phi(\mathbf{X}_2))^{\top} \rangle. \quad (1)$$

ただし、 $\mathbf{W} \in \mathbb{S}_{++}^n$ はこの距離関数のパラメータである。もし Φ が全単射なら、この距離関数 $D_{\Phi}(\cdot, \cdot; \mathbf{W}): \mathbb{S}_{++}^n \times \mathbb{S}_{++}^n \rightarrow \mathbb{R}$ は、非負性、確定性、対称性、三角不等式を満たすので、距離の公理に従うことになる。距離関数 (1) において、 $\Phi(\mathbf{X}) = \text{logm}(\mathbf{X})$ とおくと、[3] で提案されていた距離関数に等しくなる。

3 計量学習問題の定式化

本節では、距離関数 (1) におけるモデルパラメータ \mathbf{W} の学習方法を提案する。多クラス分類問題を例にとって述べる。訓練用データが $(\mathbf{X}_1, \omega_1), \dots, (\mathbf{X}_{\ell}, \omega_{\ell}) \in \mathbb{S}_+^n \times \mathbb{N}_{n_c}$ のように得られていたとする。ただし、 \mathbf{X}_i は第 i 訓練用例題の共分散記述子で、 ω_i はそのクラスラベルである。 n_c はクラス数である。これら ℓ 個の訓練用例題に対し、 K 個のペアを $(i_1, j_1), \dots, (i_K, j_K) \in \mathbb{N}_{\ell} \times \mathbb{N}_{\ell}$ を考え、各ペアに対して次の制約を与える：

$$D_{\Phi}(\mathbf{X}_{i_k}, \mathbf{X}_{j_k}; \mathbf{W}) \begin{cases} \leq b_{ub}\xi_k & \text{if } \omega_{i_k} = \omega_{j_k}, \\ \geq b_{lb}\xi_k & \text{if } \omega_{i_k} \neq \omega_{j_k}. \end{cases} \quad (2)$$

ただし、 $\xi_k = 1$ のとき、 b_{ub} および b_{lb} は、それぞれ、同じクラス間の距離の上限、異なるクラス間の距離の下限を定める定数となる。ここで、 $\omega_{i_k} = \omega_{j_k}$ なる $k \in \mathbb{N}_K$ には、 $(y_k, b_k) = (+1, b_{ub})$, $\omega_{i_k} \neq \omega_{j_k}$ なる $k \in \mathbb{N}_K$ には、 $(y_k, b_k) = (-1, b_{lb})$ とおく。この制約を満たす中で、 \mathbf{W} をなるべく単位行列に近く、 ξ_k をなるべく 1 に近くなるように最適化問題を定式化すると、

$$\min \text{BD}_{\varphi}(\mathbf{W}, \xi), (\mathbf{I}, \mathbf{1}) \quad (3)$$

$$\text{wrt } \mathbf{W} \in \mathbb{S}_{++}^n, \quad \xi = [\xi_1, \dots, \xi_K]^{\top} \in \mathbb{R}_{++}^K,$$

$$\text{subject to } \forall k \in \mathbb{N}_K, \quad y_k D_{\Phi}(\mathbf{X}_{i_k}, \mathbf{X}_{j_k}; \mathbf{W}) \leq y_k b_k \xi_k$$

のようになる。ただし、 $\text{BD}_{\varphi}(\cdot, \cdot): (\mathbb{S}_{++}^n \times \mathbb{R}_{++}^K) \times (\mathbb{S}_{++}^n \times \mathbb{R}_{++}^K) \rightarrow \mathbb{R}_+$ は、ブレグマンダイバージェンス (BD) である。 $\text{BD}_{\varphi}(\mathbf{W}, \xi), (\mathbf{I}, \mathbf{1})$ は $(\mathbf{W}, \xi) = (\mathbf{I}, \mathbf{1})$ の時に限り、0 になり、 (\mathbf{W}, ξ) が $(\mathbf{I}, \mathbf{1})$ から離れるほど値が大きくなる。 $\text{BD}_{\varphi}(\cdot, \cdot)$ は、連続微分可能で狭義凸な実数値関数 (シード関数と呼ぶ) $\varphi: \mathbb{S}_{++}^n \times \mathbb{R}_{++}^K \rightarrow \mathbb{R}$ を使って、

$$\text{BD}_{\varphi}(\Theta, \Theta_0) = \varphi(\Theta) - \varphi(\Theta_0) - \langle \nabla \varphi(\Theta_0), \Theta - \Theta_0 \rangle$$

と定義される。ただし、 $\Theta, \Theta_0 \in \mathbb{S}_{++}^n \times \mathbb{R}_{++}^K$ である。シード関数は、連続微分可能で狭義凸であれば、任意

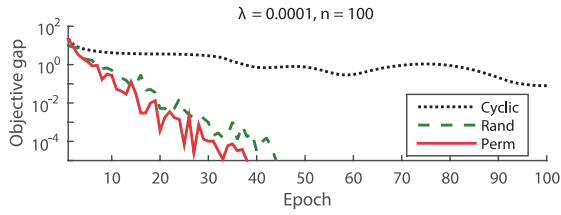


図1 最適解への収束の様子. 確率的ダイクストラ法の適用により Objective gap が急速に縮まっていることが観測される.

に定義できる. シード関数の定義によって, (\mathbf{W}, ξ) の $(\mathbf{I}, \mathbf{1})$ からの逸脱の測り方が変わり, これによって異なる学習結果を生む. 本研究では, 次のシード関数は $\varphi(\mathbf{W}, \xi) := \varphi_r(\mathbf{W}) + \sum_{k=1}^K c_k \varphi_l(\xi_k)$ の形式で与えられるとした. シード関数 φ の第1項 $\varphi_r: \mathbb{S}_{++}^n \rightarrow \mathbb{R}$ は $\varphi_r(\mathbf{W}) := -\log \det(\mathbf{W})$ とおき, 第2項の $\varphi_l: \mathbb{R}_{++} \rightarrow \mathbb{R}$ には, 次の3種類の関数を試した: $\varphi_{is}(\xi_k) := -\log(\xi_k)$, $\varphi_{l2}(\xi_k) := \xi_k^2/2$, $\varphi_e(\xi_k) := (\log \xi_k - 1)\xi_k$. 1番目のシード関数 φ_{is} から生成されるBDは, [3]で用いられた目的関数と等価になる.

4 最適化アルゴリズム

巡回的ダイクストラ法: 最適化問題(3)は, $(\mathbb{S}_{++}^n \times \mathbb{R}_{++}^K)$ 上の点 $(\mathbf{I}, \mathbf{1})$ から, K 個の半平面の共通集合へのブレッグマン射影(BP)を求める問題とみることができる. このような問題を解く方法として, ダイクストラ法[5]を利用できる. ダイクストラ法は, $(\mathbf{W}_0, \xi_0) = (\mathbf{I}, \mathbf{1})$ から開始する反復法で, 第 t 反復において, 巡回的に半空間を一つ選んで, $(\mathbf{W}_{t-1}, \xi_{t-1})$ からその半空間の境界に射影して, (\mathbf{W}_t, ξ_t) を得る. この射影を繰り返すと, 最適解に収束することが理論的に保証されている. 反復 t において, 半空間の境界 $\langle \mathbf{A}_k, \mathbf{W} \rangle = b_k \xi_k$ に射影する問題は, $\mathbf{W}_{t-1}^{-1} + \delta y_k \mathbf{A}_k > \mathbf{O}$ を満たす範囲内で, スカラー変数 δ の一変数非線形方程式

$$\langle \mathbf{A}_k, (\mathbf{W}_{t-1}^{-1} + \delta y_k \mathbf{A}_k)^{-1} \rangle = b_k \nabla \varphi_l^* (\nabla \varphi_l(\xi_{t-1}^{-1}) - \delta y_k b_k / c_k) \quad (4)$$

を解く問題に帰着される. ただし, φ_l^* は φ_l の凸共役である. 文献[3]では, スペクトラル分解や逆行列計算も不要な閉形式で解を与えてしまったが, これはウッドベリー公式の誤用であり, 実際には文献[3]の更新式ではこの方程式の解にならない. 閉形式で求められないとすると, ニュートン法などの数値算法を頼らざるを得ない. (4)の左辺には $n \times n$ 行列の逆行列を含んでいるため, ナイーブに数値算法にかけてしまうと, 計算負荷の高い

表1 BCIデータの認識率(%).

id+ \mathbf{I}	id+ \mathbf{W}	logm+ \mathbf{I}	logm+ \mathbf{W}	chol+ \mathbf{I}	chol+ \mathbf{W}
49.89	50.74	55.29	57.05	38.70	53.70

$n \times n$ 行列の逆行列を収束するまで計算し続けることになる. よって, 数値算法内部での反復回数を L とすると, 1回のBPに $O(Ln^3)$ にかかることになる. 一方, 本研究では, $O(n^3 + Ln)$ の計算量で方程式(4)の根を見つけるアルゴリズムを発見した(詳細は[4]参照).

確率的ダイクストラ法の提案: 近年, 正則化損失最小化の手法として, 確率的勾配法が注目を集めている. 確率的勾配法では, 例題を巡回的に選ぶのではなく, ランダムに選択することで, 早く最適解に収束することが実験的に確かめられている. 一方, 従来のダイクストラ法は K 個の半空間を巡回的(Cyclic)に選んでいた. 本研究では, ダイクストラ法において, K 個の半空間を巡回的に選ぶのではなく, ランダムに選んだり(Rand), 1エポックごとに K 個の半空間の順番を無作為に入れ替える(Perm)ことで, 飛躍的に最適解への収束が早くなることを発見した(図1参照).

5 脳波信号の分類実験

BCI competition IV のIIa データセットで多クラス分類のベンチマークをとった結果を表1に示す. ただし, id(恒等関数), logm, chol は Φ の種類を表す. \mathbf{I} は計量学習なし, \mathbf{W} は計量学習ありを表す. 計量学習なしの場合, 最高で55.29%であったのに対し, 計量学習ありでは, 57.05%になった. よって, 提案法により, 脳波信号の分類における汎化性能が向上することを確認した.

謝辞: 本研究はJSPS 科研費26249075, 40401236の助成を受けたものである.

参考文献

- [1] A. Barachant et al. Multiclass brain computer interface classification by riemannian geometry. *IEEE Trans Biomed Eng.*, Vol. 59, No. 4, pp. 920–928, 2012.
- [2] J. V. Davis et al. Information-theoretic metric learning. In *ICML*, pp. 209–216, New York, NY, USA, 2007. ACM.
- [3] Zhiwu Huang et al. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, pp. 720–729, 2015.
- [4] T. Matsuzawa et al. Stochastic dykstra algorithms for metric learning on positive semi-definite cone. Technical report, 2016. arXiv:1601.01422.
- [5] Y. Censor and S. Reich. The dykstra algorithm with bregman projections. *Comm. in Applied Anal.*, Vol. 2, pp. 407–419, 1998.