F-015

確率的ダイクストラ法による複数マハラノビス行列の学習
Stochastic Dykstra Algorithms for Learning Multiple Mahalanobis Matrices

大沼 由弥 † 佐藤 貴亮 † 松澤 知己 † 加藤 毅 † § ¶
Yuya Onuma Takaaki Sato Tomoki Matsuzawa Tsuyoshi Kato

1. はじめに

パターン認識問題において，特徴ベクトル間の距離計量を判別的に学習することによって識別性能が向上することは，数多くの研究報告によって裏付けられている．今日においては，対象がベクトル形式ではない場合における計量学習算法に研究の関心が拡大している [3]．本論文では，解析対象 $\boldsymbol{x} \in \mathcal{X}$ が M 個の行列の集合 $\{\boldsymbol{\Phi}_m(\boldsymbol{x})\}_{m=1}^M$ で表現される場合を考える．ただし，m 個目の行列のサイズは $n_m \times n'_m$ であり，$\boldsymbol{\Phi}_m : \mathcal{X} \to \mathbb{R}^{n_m \times n'_m}$ は対象 $\boldsymbol{x}$ から m 個目の行列を取り出す特徴抽出器である．本論文では，次の形式で表される距離関数 $D_{\boldsymbol{\Phi}}(\cdot,\cdot;\mathcal{W}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ を議論する：

$$D_{\boldsymbol{\Phi}}(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathcal{W}) := \frac{1}{M} \sum_{m=1}^{M} \Big\langle \boldsymbol{W}_m, (\boldsymbol{\Phi}_m(\boldsymbol{x}_1) - \boldsymbol{\Phi}_m(\boldsymbol{x}_2)) (\boldsymbol{\Phi}_m(\boldsymbol{x}_1) - \boldsymbol{\Phi}_m(\boldsymbol{x}_2))^\top \Big\rangle. \quad (1)$$

ただし，$\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ であり，$\mathcal{W} := (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M)$ は距離関数のパラメータで，M 個の正定値行列 $\boldsymbol{W}_1 \in \mathbb{S}^{n_1}_{++}, \ldots, \boldsymbol{W}_M \in \mathbb{S}^{n_M}_{++}$ からなる．$n = \max_m n_m$ とおく．たとえば，$M = 1$ で $\boldsymbol{\Phi}_m(\boldsymbol{x}) \in \mathbb{R}^{n \times 1}$ の場合，標準的なマハラノビス距離に等しい．$M = 1$ で $\boldsymbol{\Phi}_m(\boldsymbol{x}) \in \mathbb{S}^n_+$ が**共分散記述子**もしくはその *spectral variants* の場合に関しては，すでに幾つかの論文で議論されている (e.g. [3])．M を 2 以上にすれば，複数の共分散記述子間距離を統合した距離となる．入力空間 $\mathcal{X}$ が M モードのテンソルの集合のとき，$\boldsymbol{\Phi}_m(\boldsymbol{x})$ を mode-m flattening とすることで，テンソル間の距離関数を表すこともできる．

本論文では，上述の距離関数 (1) のパラメータ $\mathcal{W}$ の値を判別的に学習する算法を提案する．本研究では，ベクトル間の計量学習算法として有名な *ITML* [2] に倣って，**ブレグマン距離**で目的関数を構成した．ITML は，**ダイクストラ法** [1] を使って目的関数の最小化を行うものである．ダイクストラ法は，複数の**半空間**の共通集合内で最も近い点を探すための反復解法であり，各反復で半空間を巡回的に選択して解を**射影**することで解を最適解に収束させるものである．ITML では，$\boldsymbol{\Phi}_m(\boldsymbol{x})$ のランクが 1 であることから半空間への射影は閉形式で求められることを利用していた．一方，距離関数を (1) のように一般化してしまうと，半空間の射影は閉形式では求まらない．本研究では，(a) 計算量 $O(Mn^3)$ で半空間への射影を求めることができる効率的な算法を発見した．この理論的発見のみならず，本研究では次の 2 点の実験的発見を得た：(b) ダイクストラ法における半空間の選択を巡回的にではなく無作為に行うと，*Objective Gap* (最適値との差) が劇的に早く 0 に近づく；(c) 共分散行列を複数化して計量学習することでパターン認識における汎化性能が向上する．

†群馬大学大学院理工学府
§群馬大学次世代モビリティ社会実装センター (CRANTS)
¶早稲田大学規範科学総合研究所 (IIRS)

2. ブレグマン距離

本研究では，2 種類のブレグマン距離を用いる．ブレグマン距離は，**シード関数** $\varphi : \mathrm{dom}(\varphi) \to \mathbb{R}$ を通じて一般に次のように定義されている：

$$\mathrm{BD}_\varphi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_0) = \varphi(\boldsymbol{\Theta}) - \varphi(\boldsymbol{\Theta}_0) - \langle \nabla\varphi(\boldsymbol{\Theta}_0), \boldsymbol{\Theta} - \boldsymbol{\Theta}_0 \rangle,$$

ただし，シード関数は連続微分可能で狭義凸でなければならない．例えば，$\varphi((\boldsymbol{\Phi}_m(\boldsymbol{x}))_{m=1}^M) := \sum_{m=1}^M \langle \boldsymbol{W}_m, \boldsymbol{\Phi}_m(\boldsymbol{x}) \boldsymbol{\Phi}_m(\boldsymbol{x})^\top \rangle / M$ とおくと距離関数 (1) を得る．ブレグマン距離は，非対称なため，一般に**距離の公理**を満たさないが，写像 $\boldsymbol{x} \mapsto \{\boldsymbol{\Phi}_m(\boldsymbol{x})\}_{m=1}^M$ が**単射**のときは $D_{\boldsymbol{\Phi}}$ は距離の公理を満たす．$D_{\boldsymbol{\Phi}}$ は，本研究で用いる 2 種類のブレグマン距離のうち，1 種類目となる．

2 種類目は次のようなブレグマン距離である．後述する $\mathcal{W}$ の学習算法では，スラック変数 $\boldsymbol{\xi} = [\xi_k]_{k=1}^K \in \mathbb{R}^K$ を導入し，K 個の半空間

$$\mathcal{C}_k := \left\{ (\mathcal{W}, \boldsymbol{\xi}) \;\middle|\; \frac{y_k}{M} \sum_{m=1}^{M} \langle \boldsymbol{A}_{m,k}, \boldsymbol{W}_m \rangle \le y_k b_k \xi_k \right\} \quad (2)$$

の共通集合から解を探す．ただし，$(\boldsymbol{A}_{m,k}, b_k, y_k) \in \mathbb{S}^{n_m} \times \mathbb{R}_{++} \times \{\pm 1\}$ とする．後述する提案学習算法では，シード関数として，

$$\varphi(\boldsymbol{W}, \boldsymbol{\xi}) := -\frac{1}{M} \sum_{m=1}^{M} \mathrm{logdet}(\boldsymbol{W}_m) + c \sum_{k=1}^{K} \varphi_{\mathrm{l}}(\xi_k)$$

を用いた．c は正の定数である．第 2 項の $\varphi_{\mathrm{l}} : \mathbb{R}_{++} \to \mathbb{R}$ には，次の 3 種類の関数を試した：$\varphi_{\mathrm{is}}(\xi_k) := -\log(\xi_k)$, $\varphi_{\mathrm{l2}}(\xi_k) := \xi_k^2/2$, $\varphi_{\mathrm{e}}(\xi_k) := (\log \xi_k - 1)\xi_k$．1 番目のシード関数 φ_{is} から生成されるブレグマン距離は，$M = 1$, $n'_m = 1$ のとき ITML [2] の目的関数と等価になる．

3. 計量学習問題の定式化

多クラス分類問題を考える．ℓ 個の訓練用例題の集合 $\{(\boldsymbol{x}_i, \omega_i)\}_{i=1}^\ell$ が所与とする．ただし，$\boldsymbol{x}_i \in \mathcal{X}$ は入力データであり，$\omega_i \in \mathbb{N}$ はクラス番号である．これら ℓ 個の訓練用例題に対し，K 個のペアを $(i_1, j_1), \ldots, (i_K, j_K) \in \mathbb{N}_\ell \times \mathbb{N}_\ell$ を考え，各ペアに対して次の制約を与える：

$$D_{\boldsymbol{\Phi}}(\boldsymbol{x}_{i_k}, \boldsymbol{x}_{j_k}; \mathcal{W}) \begin{cases} \le b_{\mathrm{ub}} \xi_k & \text{if } \omega_{i_k} = \omega_{j_k}, \\ \ge b_{\mathrm{lb}} \xi_k & \text{if } \omega_{i_k} \ne \omega_{j_k}. \end{cases} \quad (3)$$

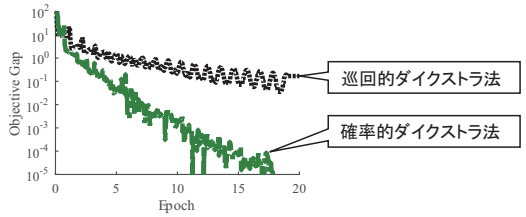

図 1: **収束の様子．** 半平面の選択を確率的にすると早く収束する．

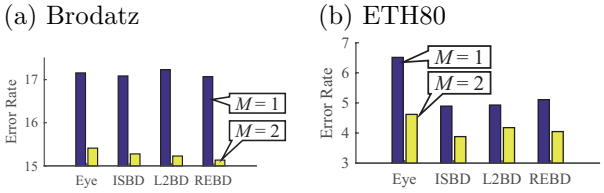

図 2: **多クラス分類の誤分類率．** Eye: $\boldsymbol{W}_m = \boldsymbol{I}$ に固定，ISBD, L2BD, REBD: 計量学習あり．それぞれ $\varphi_{\mathrm{l}} = \varphi_{\mathrm{is}}, \varphi_{\mathrm{l2}}, \varphi_{\mathrm{e}}$．

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，$(y_k, b_k) = (+1, b_{\mathrm{ub}})$，$\omega_{i_k} \neq \omega_{j_k}$ なる $k \in \mathbb{N}_K$ には，$(y_k, b_k) = (-1, b_{\mathrm{lb}})$ とおいて，$k = 1, \ldots, K$, $m = 1, \ldots, M$ に対して，

$$\boldsymbol{A}_{m,k} := \left(\boldsymbol{\Phi}_m(\boldsymbol{x}_{i_k}) - \boldsymbol{\Phi}_m(\boldsymbol{x}_{j_k})\right)\left(\boldsymbol{\Phi}_m(\boldsymbol{x}_{i_k}) - \boldsymbol{\Phi}_m(\boldsymbol{x}_{j_k})\right)^\top$$

とおくと，(3) で表される第 k 制約を満たす $(\mathcal{W}, \boldsymbol{\xi})$ の集合は半空間 $\mathcal{C}_k$ になることがわかる．

$(\mathcal{W}, \boldsymbol{\xi})$ を点 $(\mathcal{W}_0, \mathbf{1}_K)$ から共通集合 $\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_K$ への**ブレグマン射影**，すなわち，最適化問題

$$\begin{aligned} \min \quad & \mathrm{BD}_\varphi((\mathcal{W}, \boldsymbol{\xi}), (\mathcal{W}_0, \mathbf{1})), \\ \mathrm{wrt} \quad & (\mathcal{W}, \boldsymbol{\xi}) \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_K \end{aligned} \tag{4}$$

を解くことで決定することとする．ただし，$\mathcal{W}_0 := (\boldsymbol{I}_{n_1}, \ldots, \boldsymbol{I}_{n_M})$ とする．すると，マハラノビス行列はなるだけ単位行列に，また，距離制約はなるだけ $\xi_k = 1$ に近づくように $\mathcal{W}$ の値が決められることになる．

4. 確率的ダイクストラ法

射影問題 (4) をダイクストラ法で解くとすると，各反復で，K 個の半空間から一つの半空間 $\mathcal{C}_k$ を選んで，現在の解を $\mathcal{C}_k$ にブレグマン射影することになる．この射影を繰り返すことで最適解に収束することが理論的に保証されている [1]．第 t 反復における解を $(\mathcal{W}_t, \boldsymbol{\xi}_t)$，第 m マハラノビス行列を $\boldsymbol{W}_{m,t}$ とおき，$\mathcal{C}_k$ に現在の解を射影するとする．さらに，$J_t : \mathbb{R} \to \mathbb{R}$ を

$$\begin{aligned} J_t(\delta) := \sum_{m=1}^{M} \left\langle \boldsymbol{A}_{m,k}, (\boldsymbol{W}_{m,t-1}^{-1} + y_k \delta \boldsymbol{A}_{m,k})^{-1} \right\rangle \\ - M b_k \nabla\varphi_{\mathrm{l}}^*(\nabla\varphi_{\mathrm{l}}(\xi_{t-1}) + y_k \delta b_k / c), \end{aligned}$$

とおくと，$\mathcal{C}_k$ の境界への射影は各 m に対し，$\boldsymbol{W}_{m,t-1}^{-1} + \delta y_k \boldsymbol{A}_{m,k}$ が正定値である範囲で，1 変数非線形方程式 $J_t(\delta) = 0$ の根を求める問題に帰着できる．ただし，φ_{l}^* は φ_{l} の凸共役である．$J_t(\delta)$ の第 1 項には M 個の逆行列を含んでいるため，ナイーブにニュートン法などで根を求めると，1 回の射影に $O(LMn^3)$ の計算量がかかってしまう．ただし，L はニュートン法内部で関数 $J_t(\delta)$ の値を評価する回数である．本研究では，次の定理を発見した：

Theorem 1. $L \in O(n^2)$ のとき，1 回のブレグマン射影は $O(Mn^3)$ の計算量で済む．

Proof Sketch: 各 $m = 1, \ldots, M$ に対し，

$$\left\langle \boldsymbol{A}_{m,k}, (\boldsymbol{W}_{m,t-1}^{-1} + \delta y_k \boldsymbol{A}_{m,k})^{-1} \right\rangle = \sum_{i=1}^{n_m} \frac{1}{d_{i,m,k} + y_k \delta} \tag{5}$$

を満たす n_m 個の実数 $\{d_{i,m,k}\}_{i=1}^{n_m}$ を $O(n_m^3)$ で見つけることができる．いったん $\{d_{i,m,k}\}_{i=1}^{n_m}$ を見つければ，$J(\delta) = 0$ の根は $O(LMn)$ で見つけられる．1 回のブレグマン射影の計算量は $O(Mn^3)$ となる． □

半空間乱択の効果　本研究のもう一つの貢献は，半空間の選び方を巡回的にするのではなく，無作為に選ぶと劇的に早く最適解に収束することができることを経験的に発見したことである．図 1 は，$\varphi_{\mathrm{l}} = \varphi_{\mathrm{is}}$, $M = 3$, $n_1 = 10$, $n_2 = 15$, $n_3 = 20$ を用いたときの例である．この例では，巡回的に半空間を選んだ場合は，20 エポック経っても 10^{-1} 近くの Objective gap が残っているが，半空間の選択法を確率的にすると Objective gap に 10^{-5} 未満にまで到達している．

パターン認識性能　2 個の画像データセット Brodatz および ETH80 を用いて 3-近傍識別による多クラス分類のベンチマークをとった．クラス数はそれぞれ 112 および 8 である．各画像に対し，$(M = 2)$ 個の共分散記述子を得た．Brodatz では，1 個目の共分散記述子は各画素から得られる $(n_1 = 5)$ 次元ベクトルから構成，2 個目は画像解像度を半分にしたときの各画素から得られる $(n_2 = 5)$ 次元ベクトルから抽出した．ETH80 では $n_1 = n_2 = 9$ の記述子を抽出した．図 2 に，$M = 1$ の場合との誤分類率の比較を示す．計量学習により多クラス分類の汎化性能が向上することが確認できる．

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] Y. Censor and S. Reich. The dykstra algorithm with bregman projections. *Comm. in Applied Anal.*, 2:407–419, 1998.

[2] J. V. Davis et al. Information-theoretic metric learning. In *ICML*, pages 209–216, New York, NY, USA, 2007. ACM.

[3] T. Matsuzawa, R. Relator, J. Sese, and T. Kato. Stochastic dykstra algorithms for metric learning with positive definite covariance descriptors. In *ECCV2016*, pages 786–799, 2016.