7M-01

距離しきい値を自動調整できるITML型計量学習法

大沼 由弥 *　　加藤 毅 *,†,‡

* 群馬大学大学院理工学府 電子情報・数理教育プログラム

† 群馬大学次世代モビリティ社会実装研究センター　　‡ 早稲田大学規範科学総合研究所

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．本論文では，2つの入力 $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ に対するマハラノビス距離を一般化させた距離計量

$$D_{\boldsymbol{\Phi}}(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{W}) := \operatorname{tr}\left(\boldsymbol{W}\left(\boldsymbol{\Phi}(\boldsymbol{x}_1) - \boldsymbol{\Phi}(\boldsymbol{x}_2)\right)\left(\boldsymbol{\Phi}(\boldsymbol{x}_1) - \boldsymbol{\Phi}(\boldsymbol{x}_2)\right)^\top\right) \quad (1)$$

を考え，マハラノビス行列 $\boldsymbol{W} \in \mathbb{S}^n_{++}$ の教師あり学習を行う方法を議論する．ただし，$\mathbb{S}^n_{++}$ は n 次狭義正定値行列の集合である．例えば，$\boldsymbol{\Phi} : \mathcal{X} \to \mathbb{R}^n$ とすると (1) は，簡単な式変形により，標準的なベクトル間のマハラノビス距離になることがわかる．筆者らの研究グループでは $\boldsymbol{\Phi}(\boldsymbol{x})$ を $\boldsymbol{x}$ から得られる共分散記述子に一般化した場合の計量学習法を開発していた [2]．本研究では，複数のマハラノビス行列を含むように拡張した距離における計量 (1) の重みパラメータ $\mathcal{W}$ を学習するアルゴリズムを開発した．しかし簡単のため，本論文では単一のマハラノビス行列を用いる場合を中心に議論する．複数にした場合の理論の詳細は，文献 [3] に記した．

ベクトル空間における計量学習の方法としては *ITML* [1] が広く用いられており，多くの派生法が生まれている．ITML はブレグマン射影問題で学習法を構成していることから，最適化にダイクストラ法 [2] を適用できる．ITML やその派生法は距離しきい値 b_0 というハイパーパラメータがあり，b_0 の値を利用者があらかじめ決めなくてはならないという短所を持っている．その距離しきい値 b_0 が汎化性能に敏感であるため慎重に定める必要があることが，筆者らの実験で判明した．

本研究では，**(i)** ブレグマン射影問題の枠組みから逸脱することなく，距離しきい値を自動設定できる新しい計量学習法を開発した．**(ii)** 複数の共分散記述子を扱えるように距離関数を拡張した．さらに，**(iii)** ダイクストラ法の各反復で必要な半空間への射影を $O(Mn^3)$ で求められることを発見した．ただし，M はマハラノビス行列の個数である．また，**(iv)** 距離しきい値を自動設定しても，手動で距離しきい値を設定する従来の計量学習の性能と competitive であること，および，共分散記述子を複数化して計量学習を行うことでパターン認識における汎化性能が向上することを実データを使った実験により確認した．なお，理論の詳細と実験結果は，文献 [3] を参照されたい．

2 ブレグマン距離

ブレグマン距離は**シード関数** $\varphi\colon \mathrm{dom}(\varphi) \to \mathbb{R}$ を通じて

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

のように定義される．シード関数は連続微分可能かつ狭義凸であることが仮定されている．たとえば，$\varphi(\boldsymbol{\Phi}(\boldsymbol{x}))) := \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{\Phi}(\boldsymbol{x})^\top\right)$ とおくと距離関数 (1) を得る．ITML では，計量学習問題をブレグマン射影問題で定式化した．ある凸集合 $\mathcal{S}$ へのブレグマン射影とは，$\mathcal{S}$ の元のうち，ある点 $\boldsymbol{\Theta}_0$ からのブレグマン距離 $\mathrm{BD}(\cdot, \boldsymbol{\Theta}_0; \varphi)$ が最も近くなるもの，すなわち，$\arg\min_{\boldsymbol{\Theta}\in\mathcal{S}} \mathrm{BD}(\boldsymbol{\Theta}, \boldsymbol{\Theta}_0)$ である．ブレグマン距離は狭義凸関数なので，射影点は一意に定まる．

3 従来法：距離しきい値 b_0 を固定した計量学習

距離関数のパラメータ $\boldsymbol{W}$ の値を教師あり学習によって決定するために，ℓ 個のラベルありデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell \in \mathcal{X}$ を使う．ITML 型の距離計量学習は，ℓ 個の例題から K 個の例題ペアを選択する．最初の $K_+ (< K)$ 個の例題ペア $(i_k, j_k) \in \mathbb{N}^2$ $(k = 1, \ldots, K_+)$ は，同じクラスに所属する例題 $\boldsymbol{x}_{i_k}$ および $\boldsymbol{x}_{j_k}$ を選ぶ．残りの $K_-(:= K - K_+)$ 個の例題ペア $(i_k, j_k) \in \mathbb{N}^2$ $(k = K_+ + 1, \ldots, K)$ は，異なるクラスに所属する例題 $\boldsymbol{x}_{i_k}$ および $\boldsymbol{x}_{j_k}$ を選ぶ．そのうえで，$\boldsymbol{W}$ を K 個の制約

$$\begin{cases} D_{\boldsymbol{\Phi}}(\boldsymbol{x}_{i_k}, \boldsymbol{x}_{j_k}; \boldsymbol{W}) \le \xi_k, & \text{for } k = 1, \ldots, K_+, \\ D_{\boldsymbol{\Phi}}(\boldsymbol{x}_{i_k}, \boldsymbol{x}_{j_k}; \boldsymbol{W}) \ge \xi_k, & \text{for } k = K_+ + 1, \ldots, K_- \end{cases}$$

を満たす中から探す．ただし，$\boldsymbol{\xi} := [\xi_1, \ldots, \xi_K]^\top$ はスラック変数である．ITML 型の計量学習法では，$\boldsymbol{W}$ の値を決定する前に，利用者が距離しきい値 b_0 の値を決定しておく．スラック変数 $\boldsymbol{\xi}$ については，$k = 1, \ldots, K_+$ に対しては ξ_k が $b_k(b_0) := 2b_0$ から離れるほど，$k = K_-, \ldots, K$ に対しては ξ_k が $b_k(b_0) := b_0/2$ から離れるほどペナルティがかかるように目的関数が設計されている．具体的には，$\boldsymbol{b}(b_0) := [b_1(b_0), \ldots, b_K(b_0)]^\top$ とおいて，次のように目的関数を定義している：

$$P_{\mathrm{o}}(\boldsymbol{W}, \boldsymbol{\xi}\,;\, b_0) := \mathrm{BD}(\boldsymbol{W}, \boldsymbol{I}_n\,;\, \varphi_{\mathrm{r}}) + c\mathrm{BD}(\boldsymbol{\xi}, \boldsymbol{b}(b_0)\,;\, \varphi_\ell). \quad (2)$$

ただし，c は正則化パラメータである．$P_o(\cdot, \cdot\,;\, b_0)$ の第2項はスラック変数 $\boldsymbol{\xi}$ が距離しきい値 $\boldsymbol{b}(b_0)$ から離れるほどかかるペナルティで，ITML ではそのシード関数に $\varphi_\ell(\boldsymbol{\xi}) := -\sum_{k=1}^K \log(\xi_k)$ が用いられている．第1項は $\boldsymbol{W}$ の正定値性を保ち過学習を防ぐための正則化項であり，そのシード関数には $\varphi_{\mathrm{r}}(\boldsymbol{W}) := -\mathrm{logdet}(\boldsymbol{W})$ が用いられている．

ITML-type metric learning algorithm without manual tuning of distance threshold

Yuya Onuma*, Tsuyoshi Kato*,†,‡

* Education Program of Electronics and Informatics, Mathematics and Physics, Graduate School of Science and Technology, Gunma University

† Center for Research on Adoption of NextGen Transportation Systems, Gunma University

‡ Integrated Institute for Regulatory Science, Waseda University

半空間 C_k を

$$C_k := \left\{(\boldsymbol{W},\boldsymbol{\xi}) \,\middle|\, y_k D_{\boldsymbol{\Phi}}(\boldsymbol{x}_{i_k},\boldsymbol{x}_{j_k};\boldsymbol{W}) \le y_k \xi_k\right\},$$

と定義する．ただし，$k \le K_+$ には $y_k = +1$，$k > K_+$ には $y_k = -1$ とする．ITML 型学習法では，K 個の半空間 C_k の共通集合の中から $P_o(\cdot,\cdot\,;\,b_0)$ を最小にする $(\boldsymbol{W},\boldsymbol{\xi})$ を探す．$P_o(\cdot,\cdot\,;\,b_0)$ はブレグマン距離の和になっていることから 1 つのブレグマン距離になる．よって，この最適化問題はブレグマン射影問題になり，ダイクストラ法による数値的に安定した最適化が可能になる．しかし，距離しきい値 b_0 の値は，利用者が事前に定めるか，交差検証法などで決めるしかなかった．

4 提案法：距離しきい値 b_0 を自動調整する計量学習

本節では，距離しきい値 b_0 を定数ではなく，$\boldsymbol{W}$ や $\boldsymbol{\xi}$ と同時に最適化する変数として扱うことを考える．b_0 に対する正則化項を加えて，目的関数を

$$P_{\mathrm{b}0}(\boldsymbol{W},\boldsymbol{\xi},b_0) := P_{\mathrm{o}}(\boldsymbol{W},\boldsymbol{\xi}\,;\,b_0) + c_0 \mathrm{BD}(b_0,\mu_0\,;\,\varphi_{\ell 0}).$$

と定義する．ただし $c_0 > 0$ および $\mu_0 > 0$ は定数である．この目的関数で $\boldsymbol{W}$, $\boldsymbol{\xi}$ および b_0 を同時最適化する問題は，定数は増えたが，データへの依存性の高い定数は存在していない（実データを用いた実験により確認済み [3]）．しかし，$P_{\mathrm{b}0}(\boldsymbol{W},\boldsymbol{\xi},b_0)$ は，もはや一般に 1 つのブレグマン距離には変換できないためブレグマン射影問題としては扱えず，さらに，凸関数でもないために最適化が困難になってしまう．

本研究で最も大きな理論的成果は，$\boldsymbol{W}$, $\boldsymbol{\xi}$ および b_0 の同時最適化問題について，ある設定においては再びブレグマン射影問題になることを発見したことである．実行可能領域 $\bigcap_{k=1}^{K} C_k$ 内で目的関数 $P_{\mathrm{b}0}(\boldsymbol{W},\boldsymbol{\xi},b_0)$ を最小化する問題は，$P_{\mathrm{b}}(\boldsymbol{W},\boldsymbol{\xi}) := \min_{b_0 \in \mathrm{dom}(\varphi_{\ell 0})} P_{\mathrm{b}0}(\boldsymbol{W},\boldsymbol{\xi},b_0)$ を使って，

$$\min \quad P_{\mathrm{b}}(\boldsymbol{W},\boldsymbol{\xi}) \quad \mathrm{wrt} \quad (\boldsymbol{W},\boldsymbol{\xi}) \in \bigcap_{k=1}^{K} C_k \tag{3}$$

と等価的に書き換えることができる．本研究では，2 つのシード関数を

$$\varphi_{\ell 0}(b_0) := \frac{1}{2} b_0^2, \qquad \varphi_{\ell}(\boldsymbol{\xi}) := \frac{1}{2}\|\boldsymbol{\xi}\|^2 \tag{4}$$

とおいたとき，目的関数 $P_{\mathrm{b}}(\cdot,\cdot)$ は，定数を除いて 1 つのブレグマン距離になることを見つけた．

Theorem 1. $\varphi_{\ell 0}$ および φ_{ℓ} を (4) のように定義すると，任意の $\boldsymbol{\xi} \in \mathbb{R}^K$ について

$$P_{\mathrm{b}}(\boldsymbol{W},\boldsymbol{\xi}) + C = \mathrm{BD}((\boldsymbol{W},\boldsymbol{\xi}),(\boldsymbol{I},\boldsymbol{\xi}_0)\,;\,\varphi_{\mathrm{tot}}), \tag{5}$$

を満たす $\boldsymbol{G} \in \mathbb{S}^K_{++}$, $\boldsymbol{\xi}_0 \in \mathbb{R}^K$, および $C \in \mathbb{R}$ が存在する．ただし，右辺のブレグマン距離のシード関数は

$$\varphi_{\mathrm{tot}}(\boldsymbol{W},\boldsymbol{\xi}) := \varphi_{\mathrm{r}}(\boldsymbol{W}) + \frac{1}{2}\langle \boldsymbol{\xi}, \boldsymbol{G}\boldsymbol{\xi}\rangle . \tag{6}$$

とする．(証明は [3] 参照．)

複数共分散記述子への拡張　ここまでは簡単のため，距離関数を従来から用いられてきた (1) の形式のまま議論してきたが，実験により，複数の共分散記述子を抽出して次の距離を用いることでパターン認識性能が向上することが分かった：

$$D_{\boldsymbol{\Phi}}(\boldsymbol{x}_1,\boldsymbol{x}_2;\mathcal{W}) := \frac{1}{M}\sum_{m=1}^{M} \Big\langle \boldsymbol{W}_m, (\boldsymbol{\Phi}_m(\boldsymbol{x}_1) - \boldsymbol{\Phi}_m(\boldsymbol{x}_2))(\boldsymbol{\Phi}_m(\boldsymbol{x}_1) - \boldsymbol{\Phi}_m(\boldsymbol{x}_2))^{\top} \Big\rangle. \tag{7}$$

ただし，$\boldsymbol{\Phi}_m$ は第 m 特徴抽出器であり，距離関数のパラメータは M 個のマハラノビス行列 $\mathcal{W} := (\boldsymbol{W}_1,\ldots,\boldsymbol{W}_M)$ で構成している．Theorem 1 は (7) を使用した場合にも容易に拡張できる [3].

最適化アルゴリズム　Theorem 1 に示したとおり，最適化問題 (3) はブレグマン射影になるので，ダイクストラ法を使用することができる．確率的ダイクストラ法は反復法のひとつで，反復 $t-1$ において，解が $(\mathcal{W}_{t-1},\boldsymbol{\xi}_{t-1})$ にいるとすると，反復 t では無作為に選んだ第 k 半空間の境界に解 $(\mathcal{W}_{t-1},\boldsymbol{\xi}_{t-1})$ を射影する．その射影点は次の $\bar{\delta}$ の非線形一元方程式

$$\langle e_k, \boldsymbol{\xi}_{t-1}\rangle + \bar{\delta} h_{k,k} = \frac{1}{M}\sum_{m=1}^{M} \left\langle \boldsymbol{A}_{m,k}, \left(\boldsymbol{W}_{t-1,m}^{-1} + \bar{\delta}\boldsymbol{A}_{m,k}\right)^{-1} \right\rangle \tag{8}$$

を解くと得られる [3]．ただし，

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

$h_{k,k}$ は $\boldsymbol{G}^{-1}$ の第 k 対角成分である．非線形方程式 (8) の解は閉形式では求まらないので，二分探索やニュートン法などを使って数値的に探さなくてはならない．ナイーブに探そうとすると，いくつかの $\bar{\delta}$ の値における (8) の両辺を評価することになる．数値解法内部で L 回両辺を評価すると，(8) の右辺は計算に $O(n^3)$ かかる逆行列を M 個含んでいるため，この非線形方程式の解を見つけるのに $O(LMn^3)$ の計算量がかかる．本研究では，文献 [2] で用いたトリックを再利用することで，$O(Mn^3)$ の計算量で非線形方程式を解くことができることを見出した [3].

実験結果　実験結果は文献 [3] にて報告する．

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

参考文献

[1] Davis J. V. et al.: Information-theoretic metric learning, *ICML*, ACM, pp. 209–216 (2007).
[2] Matsuzawa, T., Relator, R., Sese, J. and Kato, T.: Stochastic Dykstra Algorithms for Metric Learning with Positive Definite Covariance Descriptors, *ECCV*, pp. 786–799 (2016).
[3] Onuma, Y., Rivero, R. and Kato, T.: Threshold Auto-Tuning Metric Learning, `https://arxiv.org/abs/1801.02125`.