7ZF-03

# 符号制約学習によるタンパク質機能予測

下山　愛祐美 * 小林　美里 * 加藤 毅 * ‡ §
Ayumi Shimoyama Misato Kobayashi Tsuyoshi Kato

## 1. はじめに

分子生物学の分野では，個々のタンパク質の機能を明らかにすることが細胞内のメカニズムを解明するための重要なステップとされている．機能未知のタンパク質の機能を予測するために配列類似度は重要な手段である．配列類似度と判別的学習を組み合わせることで予測精度を向上させる **SVM-Pairwise 法** [2] が提案され，それ以来多くの研究報告においてこのアプローチの有効性が実証されてきた (e.g. [1])．SVM-Pairwise 法では，$n$ 個のアミノ酸配列を訓練用に使うことができるとき，特徴ベクトルを $n$ 次元で構成し (i.e. $\boldsymbol{x} = [x_1, \ldots, x_n]^\top$)，$n$ 個の特徴はそれぞれ訓練用例題との配列類似度としている．$n$ 個の配列のうち，最初の $n_+$ 個のタンパク質が正例，残りが負例とすると，$n_+$ 個の特徴 $x_1, \ldots, x_{n_+}$ は正例との配列類似度になり，$n_-(:= n - n_+)$ 個の特徴 $x_{n_+ +1}, \ldots, x_n$ は負例との配列類似度となる．このように構成した特徴ベクトルを SVM に与え，重み係数 $\boldsymbol{w} := [w_1, \ldots, w_n]^\top$ の値を決定する．そのうえで，ターゲットとなるタンパク質の予測スコアを

$$\sum_{i=1}^{n_+} w_i x_i + \sum_{i'=n_+ +1}^{n} w_{i'} x_{i'} \qquad (1)$$

によって求める．この予測スコアが一定以上の値なら入力配列は特定の機能を持つと予測する．(1) からわかるように，$n_+$ 個の重み係数 $w_1, \ldots, w_{n_+}$ の符号は非負であり，$n_-$ 個の重み係数 $w_{n_+ +1}, \ldots, w_n$ の符号は非正であることが望ましいが，SVM-Pairwise 法では重み係数の符号がこのようになることは保証されない．本研究では，重み係数に

$$w_1 \geq 0, \ldots, w_{n_+} \geq 0, w_{n_+ +1} \leq 0, \ldots, w_n \leq 0 \qquad (2)$$

という制約を明示的につけて学習することを考案した．(2) を**符号制約**と呼ぶことにする．

本研究の成果は，以下の通りである：

- SVM および**ロジスティック回帰** (**LR**) に符号制約を導入して学習するアルゴリズムを提案する (但し，SVM に関しては文献 [4] で報告済み)．
- 開発した符号制約学習のための最適化アルゴリズムの反復数が符号制約なしの反復数と同等の回数で抑えられることを理論的に示す．
- 配列類似度による判別的学習に符号制約学習を適用することを新たに考案した．実データを使った実験結果を通して有用性を示す．

*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研究センター (CRANTS)
§早稲田大学規範科学総合研究所 (IIRS)

## 2. 符号制約リスク最小化問題

**SVM** や **LR** など多くの機械学習では，**正則化経験リスク**の最小化問題を解いている．いま，線形識別器 (1) の重み係数 $\boldsymbol{w} \in \mathbb{R}^n$ の値を決定するために，$n$ 個のアミノ酸配列から得られる訓練用例題 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^n \times \{\pm 1\}$ を収集したとする．すると，正則化経験リスクは

$$P(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \phi(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \qquad (3)$$

で与えられる．ただし，$\phi : \mathbb{R} \to \mathbb{R}$ は**凸上限サロゲート損失**であり，**ヒンジ損失** $\phi(z) := \max(0, 1 - z)$ に設定すると SVM になり，**ロジスティック損失** $\phi(z) := \log(1 + \exp(-z))$ に設定すると，LR になる．本研究では，この 2 つの損失関数を想定して，符号制約 (2) を課して正則化経験リスク $P(\boldsymbol{w})$ の最小化を考案し，重み係数 $\boldsymbol{w}$ の値を決定する方法を開発した．

## 3. 最適化アルゴリズム

本研究では，符号制約 (2) の下で $P(\boldsymbol{w})$ を最小化するために，**双対問題**を**確率的座標上昇法**で解く．

**Lemma 3.1.** 以下で定義される $D(\boldsymbol{\alpha})$ の最大化問題は符号制約 (2) の下で $P(\boldsymbol{w})$ を最小化する問題の双対問題である：

$$D(\boldsymbol{\alpha}) := -\frac{\lambda}{2}\|\boldsymbol{\pi}(\bar{\boldsymbol{w}}(\boldsymbol{\alpha}))\|^2 - \frac{1}{n}\sum_{i=1}^{n} \phi^*(-\alpha_i). \quad (4)$$

ただし，$\bar{\boldsymbol{w}}(\boldsymbol{\alpha}) := \sum_{i=1}^{n} y_i \boldsymbol{x}_i \alpha_i / (\lambda n)$，$\boldsymbol{\pi}(\bar{\boldsymbol{w}}) := \boldsymbol{c} \odot \max(0, \boldsymbol{c} \odot \bar{\boldsymbol{w}})$，$\boldsymbol{c} := \left[\mathbf{1}_{n_+}^\top, -\mathbf{1}_{n_-}^\top\right]^\top$ とした．$\phi^*$ は $\phi$ の**凸共役**である．

確率的座標上昇法では，各反復において，**(i)** 双対変数 $\boldsymbol{\alpha} \in \mathbb{R}^n$ に含まれる $n$ 個の変数の一つ $\alpha_i$ を無作為に選ぶ．**(ii)** 次に，

$$J_t^0(\Delta\alpha) := D(\boldsymbol{\alpha}^{(t-1)} + \Delta\alpha \boldsymbol{e}_i) - D(\boldsymbol{\alpha}^{(t-1)}) \geq 0$$

なる $\Delta\alpha$ を探す．**(iii)** 最後に $\boldsymbol{\alpha}^{(t)} := \boldsymbol{\alpha}^{(t-1)} + \Delta\alpha \boldsymbol{e}_i$ と更新する方法である．ただし，上付き添え字は反復番号を示す．$J_t^0(\Delta\alpha)$ を最大化する $\Delta\alpha$ を選ぶのが理想的であるが，一般に閉形式で最適な $\Delta\alpha$ の値は求まらない．代わりに，$J_t^0(\Delta\alpha)$ の下限を導入し，下限を最大化する $\Delta\alpha$ を選ぶ方法を考える．本研究では，$J_t^0(\Delta\alpha)$

の下限として，

$$J_t^1(\Delta\alpha) := -\frac{\|\boldsymbol{x}_i\|^2}{2\lambda n^2}(\Delta\alpha)^2 - \frac{z_i\Delta\alpha}{n} + \frac{1}{n}\left(\phi^*(-\alpha_i^{(t-1)}) - \phi^*(-\alpha_i^{(t-1)} - \Delta\alpha)\right)$$

を導入する．ただし，$z_i := y_i \left\langle \boldsymbol{x}_i, \boldsymbol{\pi}(\bar{\boldsymbol{w}}(\boldsymbol{\alpha}^{(t-1)}))\right\rangle$ とする．不等式 $J_t^0(\Delta\alpha) \geq J_t^1(\Delta\alpha)$ は Lemma 3.1 において導入した演算子 $\boldsymbol{\pi}(\cdot)$ の性質

$$\forall \boldsymbol{v}, \forall \boldsymbol{\delta} \in \mathbb{R}^d,\ \|\boldsymbol{\pi}(\boldsymbol{v}) + \boldsymbol{\delta}\| \geq \|\boldsymbol{\pi}(\boldsymbol{v} + \boldsymbol{\delta})\| \qquad (5)$$

を使って発見した．著者らの研究グループでは，損失関数 $\phi$ をヒンジ損失としたとき，$\operatorname{argmax}_{\Delta\alpha} J_t^1(\Delta\alpha)$ が閉形式で与えられることを，すでに文献 [4] にて報告している．しかし，ロジスティック損失では，依然 $\operatorname{argmax}_{\Delta\alpha} J_t^1(\Delta\alpha)$ は閉形式で求まらない．そこで，本研究では，さらに，$J_t^1(\Delta\alpha)$ の下限

$$J_t^2(\Delta\alpha) := -\frac{2(\Delta\alpha)^2}{n\bar{s}_i} + \frac{\Delta\alpha}{nq}\left(F_t + 2q^2\right) \qquad (6)$$

を導入する．ただし，$F_t := \phi(z_i) + \phi^*(-\alpha_i^{(t-1)}) + \alpha_i^{(t-1)} z_i$，$q := -\nabla\phi(z_i) - \alpha_i^{(t-1)}$，$\bar{s}_i := 4\lambda n/(4\lambda n + \|\boldsymbol{x}_i\|^2)$ とする．不等式 $J_t^1(\Delta\alpha) \geq J_t^2(\Delta\alpha)$ は，損失関数がロジスティック損失のとき，閉区間 $I_q$ で成立する．ただし，$q \geq 0$ では $I_q := [0, q]$，$q < 0$ なら $I_q := [q, 0]$ とする．$\Delta\alpha$ は区間 $I_q$ 内で $J_t^2(\Delta\alpha)$ を最大化する値に定めることにする．そのような $\Delta\alpha$ は

$$\Delta\alpha = \mathrm{Clip}_{I_q}\left[\frac{\bar{s}_i}{4q}\left(F_t + 2q^2\right)\right] \qquad (7)$$

と閉形式で表すことができる．ただし，Clip は $\mathrm{Clip}_{[a,b]}(x) = \max(a, \min(b, x))$ となる演算子である．主変数 $\boldsymbol{w}^{(t)}$ は得られた $\boldsymbol{\alpha}^{(t)}$ から，$\boldsymbol{w}^{(t)} := \boldsymbol{\pi}(\bar{\boldsymbol{w}}(\boldsymbol{\alpha}^{(t)}))$ によって復元することができる．

**収束解析：** 本研究で最も大きな理論的発見は, 主変数 $\boldsymbol{w}^{(t)}$ が $\epsilon$ 最適解に達するまでの反復数について，上述の 符号制約 LR のための学習アルゴリズムと符号制約なしの正則化リスク最小化のための SDCA 法 [3] とで等しくなることである．

**Theorem 1.** 最適解を $\boldsymbol{w}^\star \in \mathbb{R}^d$ で表すとする．損失関数にロジスティック損失を用いた上述の学習アルゴリズムは，

$$t \geq \frac{1}{s_1}\log\left(\frac{1}{\epsilon s_1}\right) \qquad (8)$$

を満たす反復 $t$ において $\mathbb{E}[P(\boldsymbol{w}^{(t)}) - P(\boldsymbol{w}^\star)] \leq \epsilon$ を保証する．ただし，$R := \max_i \|\boldsymbol{x}_i\|$，$s_1 := 4\lambda/(4\lambda n + R^2)$ とする．

表 1: タンパク質機能分類の ROC スコア.

| カテゴリ | SF-LR | SC-LR | SF-SVM | SC-SVM |
|---|---|---|---|---|
| 1 | 0.649 | **0.793** | 0.669 | 0.790 |
| 2 | 0.599 | 0.762 | 0.594 | **0.769** |
| 3 | 0.656 | **0.755** | 0.666 | 0.754 |
| 4 | 0.692 | **0.790** | 0.712 | 0.787 |
| 5 | 0.649 | 0.823 | 0.654 | **0.828** |
| 6 | 0.590 | **0.733** | 0.599 | 0.732 |
| 7 | 0.626 | 0.755 | 0.636 | **0.756** |
| 8 | 0.584 | **0.720** | 0.575 | 0.717 |
| 9 | 0.541 | 0.680 | 0.534 | **0.691** |
| 10 | 0.639 | 0.721 | 0.643 | **0.724** |
| 11 | 0.554 | 0.558 | 0.548 | **0.561** |
| 12 | 0.830 | 0.920 | 0.836 | **0.923** |

**Proof Sketch.** 各反復 $t$ における例題の選択で期待値をとると，

$$\mathbb{E}\left[J_0^2(\Delta\alpha)\right] \geq s_1^{-1}\left(P(\boldsymbol{w}^{(t-1)}) - D(\boldsymbol{\alpha}^{(t-1)})\right) \qquad (9)$$

を得ることができる．この下限は，符号制約のない SDCA 法における下限に等しい．ここに，文献 [3] で用いられている証明技法を適用すると，題意は示される． □

## 4. 実験結果

提案する符号制約 SVM (**SC-SVM**) および符号制約 LR(**SC-LR**) を出芽酵母におけるタンパク質の機能予測問題に適用した．用いたデータセットは文献 [1] と同一で，3,583 個のアミノ酸配列を含み，12 種類の機能カテゴリに分類されている．複数のカテゴリを無作為に 50% を訓練用，残りを評価用とした．正則化パラメータ $\lambda$ は訓練用データ内で交差確認法を行うことで $\{10^{-2}, 10^{-1}, 10^{0}\}$ から選択した．SC-SVM や SC-LR に加えて，従来 SVM(**SF-SVM**) や従来 LR(**SF-LR**) にもこの手続きを 5 回繰り返して，各手法，各機能カテゴリに対して，5 個の ROC スコアを得た．5 個の ROC スコアの平均を表 1 に示す．太字は 4 手法で最高の値を示し，下線は最高値と統計的有意差がないことを表す．ただし，有意水準 1%で 1 標本 t 検定を用いた．その結果，ほとんどの機能クラスに対して，符号制約学習によって統計的有意に予測精度の向上が確認された．SVM と LR との間には大差はみられなかった．

**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] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, -(-):300–11, - 2004.

[2] L. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J Comput Biol*, 10(6):857–68, dec 2003.

[3] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.*, 14(1):567–599, February 2013.

[4] 小林 美里, 佐野 大輔, 加藤 毅. 水文水質データを利用した大腸菌予測のための符号制限学習. In *第 16 回情報科学技術フォーラム FIT2017*, 第 1 分冊, pages 103–104, Sept 2017.