A-007

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
Learning Sign-Restricted Linear Classifier for Prediction of *Escherichia coli* Counts in River Water

小林 美里 † 佐野 大輔 ‡ 加藤 毅 † § ¶
Misato Kobayashi Daisuke Sano Tsuyoshi Kato

1. まえがき

線形予測器 $\langle \boldsymbol{w}, \cdot \rangle$ のモデルパラメータ $\boldsymbol{w} := [w_1, \ldots, w_d]^\top \in \mathbb{R}^d$ の学習タスクにおいて，高い汎化能力を得るには，十分な個数の訓練用データが必要である．しかし，応用によっては，十分な訓練用データが得られない場合がしばしばある．特に，医学や生物学などにおいて，1 個のデータ点を得るのに，高価な試薬や，少なからぬ労力を要するような応用は珍しくない．訓練用データの個数が不十分でも，訓練の精度を向上させるための有効な方法としては，事前知識の活用が知られている．

応用分野によっては，ドメイン知識として，ある説明変量 x_h が出力変量 y と正の相関があることが分かっているような場合がある．そのような場合，訓練用データの個数が十分にあれば，対応するモデルパラメータ w_h は正になると予想される．しかし，訓練用データの個数が，次元数に比べて十分ではないようなときや，クラス間の分布の重なりが大きいとき，対応するモデルパラメータ w_h が負に学習されてしまうこともしばしば起こり，その結果，その説明変量が正しい予測を妨げてしまう．

本論文では，一部のモデルパラメータの符号があらかじめ分かっているときに，そのドメイン知識を組み込む学習アルゴリズムを提案する．本研究では，SVM 学習に符号の制約を加えた最適化問題を扱う．標準的な SVM 学習問題の最適化法のうち，今日もっとも主流になっている手法は確率的勾配法 (SGD) と SDCA 法 [1] の 2 つである．SDCA 法はほかの最適化手法と比べて次の長所がある．(i) SDCA 法はステップサイズの設定が不要．SGD 法は，手動でステップサイズを設定しなくてはならず，この値が小さすぎても大きすぎても最適解に到達しない．(ii) SGD 法は，停止条件がはっきりしておらず，結局，一定数，もしくは，ほとんど解が更新しなくなったら，といった条件を使わざるを得ない．SDCA 法は，最小値との差の上限を計算することができるので，算法の停止条件を明確に定義でき，算法を停止させた時の最適性を保証できる．(iii) SGD 法は，弱い正則化定数 λ を用いたとき，SDCA 法と比べて，最適解への収束が遅い．

本研究では，SDCA 法をベースにして，SVM のパラメータに符号制約を加えた場合の最適化算法を開発する．制約空間内で最適化するときの典型的な方法として，勾配射影法がある．勾配射影法は，勾配法の各勾配において，制約空間へ射影するステップを加えたものである．本研究で開発した算法は，SDCA 法の各反復において制約空間へ射影するステップを加えたものである．提案法は，このステップを挿入したにも関わらず，SDCA 法の長所がほとんど失われていない．特に，*SDCA* **法に制約空間に射影するステップを加えたとしても，*SDCA* 法の収束率が変化しないという理論的結果**を得た．

このたびの成果は，これまでの我々の研究結果 [3] をオーダーレベルで改善するものでもある．文献 [3] で発表した SGD 法ベースの算法の収束率は $O(1/(\lambda\epsilon))$ であるが，**本論文で提案する算法は** $O((n+(1/\lambda))\log(1/\epsilon))$ **の収束率に改善**した．

この算法で学習した識別器を河川の水文水質データから大腸菌数を予測する問題に適用した．水文水質データからなる各説明変量の性質は分かっているので，その知識に基づいて SVM のパラメータに符号の制限をかけて学習した．加えて，タンパク質機能予測問題にも適用した．この 2 つのタスクにおいて，符号制限をつけて学習した識別器は，汎化性能を飛躍的に向上させることができた．

2. 提案法：符号制限 SVM の学習算法

SVM の識別関数は，説明変量 $\boldsymbol{x} \in \mathbb{R}^d$ に対して，$\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ で与えられる．SVM 学習では，サイズ n の訓練用データセット $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ $(i = 1, \ldots, n)$ から定義される目的関数

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

を最小化する $\boldsymbol{w}$ を見つける．ただし，λ は正則化パラメータであり，$l(\cdot)$ は損失関数であり，

$$l(z) := \begin{cases} 1 - z - m/2 & \text{if } z \in (-\infty, 1-m], \\ \frac{1}{2m}(z-1)^2 & \text{if } z \in (1-m, 1], \\ 0 & \text{if } z \in (1, +\infty) \end{cases}$$

で与えられる．これは**スムース化ヒンジ損失**と呼ばれる．通常よく用いられてきた標準的なヒンジ損失はスムース化ヒンジ損失において $m \to 0$ としたものに等しい．SDCA 法では，この最小化問題の Fenchel 双対を考える．その双対変数を $\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_n]^\top \in \mathbb{R}^n$ とする．反復 t において，n 個の双対変数のうち，一つだけ無作為に選ぶ．選んだ変数を α_i とすると，反復 t においてその変数は

$$\alpha_i \leftarrow \max\left(0, \min\left(1, \frac{1 - y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - \|\boldsymbol{x}_i\|^2 \alpha_i^{\mathrm{old}}/(\lambda n)}{\|\boldsymbol{x}_i\|^2/(\lambda n) + m}\right)\right) \tag{1}$$

†群馬大学大学院理工学府
‡東北大学工学研究科
§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)

のように更新されるものであった [1].

従来の SVM では，$\mathbb{R}^d$ 全体から $f(\boldsymbol{w})$ を最小化する．提案法 (**符号制限 SVM**) では，特定の次元の符号を制限する．すなわち，ドメイン知識を使って，添え字集合 $\{1,\ldots,d\}$ の部分集合 $\mathcal{I}_+$ および $\mathcal{I}_-$ を指定して，モデルパラメータ $\boldsymbol{w}$ を制約集合

$$\mathcal{S} := \left\{\boldsymbol{w} \in \mathbb{R}^d \,|\, \forall h \in \mathcal{I}_+, w_h \geq 0, \forall h' \in \mathcal{I}_-, w_{h'} \leq 0\right\}$$

の中から探すこととする．本研究で開発した最適化算

Algorithm 1 符号制限 SVM の学習算法

```
begin
α := 0_n; w̄ := 0_d; w := 0_d;
for t = 1, 2, ... do
    Pick i randomly from {1, ..., n};
    α_i^old ← α_i;
    Update α_i using (1).
    w̄ ← w̄ + (α_i − α_i^old) y_i x_i/(λn);
    w := Proj_S(w̄);
end for
end.
```

法は，Algorithm 1 に示すように，通常の SDCA 法に射影ステップ (ステップ 8) を挿入したものになっている．ただし，$\mathrm{Proj}_{\mathcal{S}}(\bar{\boldsymbol{w}})$ は，点 $\bar{\boldsymbol{w}}$ から制約空間 $\mathcal{S}$ への射影を表す．この射影は，次元ごとに，$h \in \mathcal{I}_+$ に対して $w_h \leftarrow \max(0, \bar{w}_h)$，$h' \in \mathcal{I}_-$ に対して $w_{h'} \leftarrow \min(0, \bar{w}_{h'})$ を行うことに等しい．すなわち，符号制約に違反した次元の値を単純に 0 に戻す変換になる．この射影ステップの導入によって，各反復が終了した時点では $\boldsymbol{w} \in \mathcal{S}$ が成立する．

Algorithm 1 に対して，本研究では次のことを発見した．

Theorem 1. 最適解を $\boldsymbol{w}^\star \in \mathbb{R}^d$ で，反復 t における解を $\boldsymbol{w}^{(t)} \in \mathbb{R}^d$ で表すとする．Algorithm 1 は次を満たす反復 t において $\mathbb{E}[f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^\star)] \leq \epsilon$ を満たす：

$$t \geq \left(n + \frac{1}{\lambda m}\right) \log\left(\frac{1}{\epsilon}\left(n + \frac{1}{\lambda m}\right)\right). \quad (2)$$

すなわち，各反復において符号を強制的に修正するステップを入れたとしても，従来の SVM のための SDCA 法の収束率 $O((n+\frac{1}{\lambda})\log(\frac{1}{\epsilon}))$ を維持するという理論的結果を得た．

Theorem 1 の Proof Sketch. SVM において，正則化関数を $\|\boldsymbol{w}\|^2/2$ から，$g(\boldsymbol{w}) = \|\boldsymbol{w}\|^2/2 + \delta_{\mathcal{S}}(\boldsymbol{w})$ に変更することを考える．ただし，$\delta_{\mathcal{S}}(\boldsymbol{w})$ は集合 $\mathcal{S}$ への指示関数であり，$\boldsymbol{w} \in \mathcal{S}$ では $\delta_{\mathcal{S}}(\boldsymbol{w}) = 0$，$\boldsymbol{w} \notin \mathcal{S}$ では $\delta_{\mathcal{S}}(\boldsymbol{w}) = +\infty$ となる．この正則化関数 g は強凸であるため，Prox-SDCA 法 [2] の仮定を満たす．また，g の凸共役の導関数 $\nabla g^*(\bar{\boldsymbol{w}})$ は，$\mathcal{S}$ への射影となることを示すことができ，すると，Algorithm 1 が導かれる．Prox-SDCA 法と SDCA 法は収束率が等しい [2] ので，Theorem 1 は成立する． □

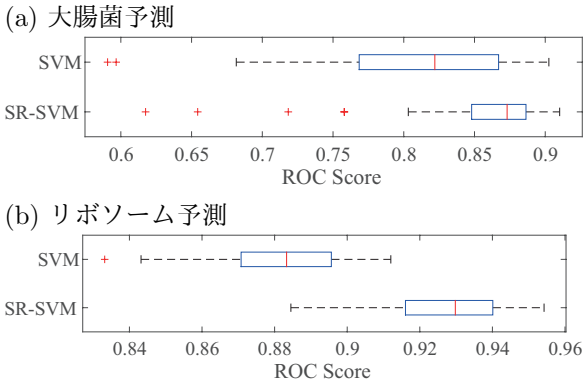

図 1: 汎化性能の比較．

3. 汎化性能の評価実験

提案する符号制限 SVM を，河川の水文水質データから大腸菌数を予測するタスクに適用した．水環境工学で培われた知見により，説明変量 WT, EC, SS, BOD, TN, TP の係数 w_h を非負，説明変量 pH_+, pH_-, DO, 流量の係数 w_h は非正に制限して学習を行った．実験条件の詳細は文献 [3] を参照のこと．また，アミノ酸配列から各タンパク質が，リボソームか否か，予測する問題にも適用した．訓練用データと評価用データを無作為に分割して，ROC スコアで汎化性能を評価した．これを 50 回繰り返し，箱ひげ図でプロットしたのが，図 1 である．符号制約の効果が明白に表れており，提案法は強力なアプローチであることが実証された．

4. おわりに

本論文では，SVM の各パラメータの符号を制限することでドメイン知識を導入して学習する符号制限 SVM 法を提案した．また，SDCA 法 をベースにした最適化算法は，制約を加えても収束率が悪化しないことを証明した．これを河川大腸菌予測，およびタンパク質機能予測問題に応用し，訓練用データが小さくても高い汎化性能を維持できることを実証した．

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

[2] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2016.

[3] 小林 美里, 宮村 明帆, 佐野 大輔, and 加藤 毅. 符号制限線形識別器の開発と河川水中大腸菌数予測への応用. In 第 *15* 回情報科学技術フォーラム *FIT2016*, 第 1 分冊, pages 149–150, Sept 2016.