A-022

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
Development of Sign-Restricted Linear Classifier and Its Application to Prediction of *Escherichia coli* Counts in River Water

小林 美里 † 宮村 明帆 ‡ 佐野 大輔 ‡ 加藤 毅 †
Misato Kobayashi Akiho Miyamura 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，いつも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 は正になると期待される．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，データが noisy なとき，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，確率勾配法 (SGD) が主流となっている．SGD は，主目的関数を n 個の項に分解して，各反復で，項を一部だけ無作為に選択して，勾配と逆方向に解を移動させる．本研究では，SGD の一種である **Pegasos 法** [2] に注目する．λ を正則化パラメータ (式 (1)) とすると，Pegasos 法は，任意の $\delta \in (0,1)$ に対して，$O(1/(\lambda\epsilon\delta))$ 回の反復回数で ϵ 最適解を得る確率が $1-\delta$ 以上になることが理論的に保証されている．

本研究では，Pegasos 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した．提案する算法は，SDCA 法の各反復において制約空間に射影するステップを加えたものである．2 節にて，この射影ステップを挿入しても，Pegasos 法の収束率が変化しない，という理論的結果を示す．

さらに，河川における大腸菌数を水門水質データから予測する問題に適用した結果を 3 節にて示す．大腸菌数のリアルタイムモニタリングが実現すれば，河川における微生物学的安全性を保持するための有用なツールとなる．しかし，分子生物学的モニタリングをリアルタイムで行うのは技術的に不可能なのが現状である [3]．

Algorithm 1 符号制限 SVM の学習算法

begin
$\boldsymbol{w} := \mathbf{0}_d$;
for $t = 1, 2, \ldots$ **do**
  Pick a subset $\mathcal{A}_t \subseteq \mathbb{N}_m$ randomly with $|\mathcal{A}_t| = k$;
  $\boldsymbol{w} := (1 - t^{-1})\boldsymbol{w} + (\lambda t k)^{-1} \sum_{i \in \mathcal{A}_t^+} \boldsymbol{x}_i y_i$;
  Project $\boldsymbol{w}$ onto $\mathcal{S}$;
  $\boldsymbol{w} := \min\{1, 1/(\sqrt{\lambda}\|\boldsymbol{w}\|)\}\boldsymbol{w}$;
end for
end.

水門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ところ，汎化性能を飛躍的に向上させることに成功したことを報告する．

2. 符号制限 SVM の提案

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 $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ $(i = 1, \ldots, m)$ から定義される目的関数

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

を最小化する $\boldsymbol{w}$ を見つける．

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法は，Algorithm 1 に示すように，従来の Pegasos 法 [2] に射影ステップ (ステップ 6) を挿入したものになっている．ただし，$\mathcal{A}_t^+ := \{i \in \mathbb{N}_m \,|\, 1 > y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\}$，$\mathbb{N}_m := \{i \in \mathbb{N} \,|\, i \leq m\}$ とする．制約空間 $\mathcal{S}$ への射影は，次元ごとに，$h \in \mathcal{I}_+$ に対して $w_h \leftarrow \max(0, w_h)$，$h' \in \mathcal{I}_-$ に対して $w_{h'} \leftarrow \min(0, w_{h'})$ 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．

Algorithm 1 に対して，本研究では次の発見をした．

†群馬大学理工学府
‡北海道大学工学研究院

Theorem 1. $\forall i, \|\boldsymbol{x}_i\| \leq R$ とする．$\boldsymbol{w}^\star := \operatorname{argmin}_{\boldsymbol{w}\in\mathcal{S}} f(\boldsymbol{w})$ とし，反復 t 直後における解を $\boldsymbol{w}^{(t)} \in \mathbb{R}^d$ であらわすとする．Algorithm 1 において各反復で得られる T 個の解 $\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)}$ から無作為に一つ選んだものを $\boldsymbol{w}^{(r)}$ とする．このとき，任意の $\delta \in (0,1)$ に対して，

$$f(\boldsymbol{w}^{(r)}) \leq f(\boldsymbol{w}^\star) + \frac{(\sqrt{\lambda}+R)^2 \log(T)}{\lambda T \delta}$$

を満たす確率が $1-\delta$ 以上になる．

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，Pegasos 法の収束率を維持するという理論的結果を得た．

Proof Sketch: Pegasos 法の収束率の証明 [2] と同様，次の補題を用いる．

Lemma 2.1. [1] $f_1, \ldots, f_T$ を λ 強凸関数とする．$\mathcal{C}$ を閉凸集合とする．$\Pi_{\mathcal{C}}(\boldsymbol{w}) := \operatorname{argmin}_{\boldsymbol{w}'\in\mathcal{C}} \|\boldsymbol{w}' - \boldsymbol{w}\|$ とおく．$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{T+1}$ を $\boldsymbol{w}_1 \in \mathcal{C}$ とし，$t \geq 1$ に対して，$\boldsymbol{w}_{t+1} := \Pi_{\mathcal{C}}(\boldsymbol{w}_t - \nabla_t/(\lambda t))$ とする．ただし，$\nabla_t \in \partial f_t(\boldsymbol{w}_t)$ である．$\forall t \in \mathbb{N}$, $\|\nabla_t\| \leq G$ を仮定する．このとき，$\forall \boldsymbol{u} \in \mathcal{C}$ に対して，以下が成立する：

$$\frac{1}{T}\sum_{t=1}^{T} f_t(\boldsymbol{w}_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(\boldsymbol{u}) + \frac{(1+\log(T))G^2}{2\lambda T}.$$

Algorithm 1 がこの補題の条件を満たすことをしめす．まず，$\mathcal{C} := \mathcal{B} \cap \mathcal{S}$ と設定する．ただし，$\mathcal{B} := \{\boldsymbol{w} \in \mathbb{R}^d \mid \|\boldsymbol{w}\| \leq \lambda^{-1/2}\}$ とする．すると，Algorithm 1 から生成される解は $\boldsymbol{w}^{(t+1)} = \Pi_{\mathcal{C}}(\boldsymbol{w}^{(t)} - \nabla_t/(\lambda t))$ を満たすことを示すことができる．さらに，$\boldsymbol{w}^\star \in \mathcal{C}$ であることは次のように示せる．双対変数 $\boldsymbol{\alpha} \in [0,1]^m$ を導入し，$\boldsymbol{v}(\boldsymbol{\alpha}) = (m\lambda)^{-1}\boldsymbol{X}\boldsymbol{\alpha}$ とおく．ただし，$\boldsymbol{X}$ は $d \times m$ 行列で第 i 列は $y_i\boldsymbol{x}_i$ とする．関数 $\boldsymbol{w} \mapsto f(\boldsymbol{w}) + \delta_{\mathcal{S}}(\boldsymbol{w})$ を最小化する問題の Fenchel 双対を取ったときに得られる目的関数を $g : \mathbb{R}^m \to \mathbb{R}$，その最適解を $\boldsymbol{\alpha}^\star$ として，強双対定理を適用すると，

$$\begin{aligned}
&\|\boldsymbol{w}^\star\|^2/2 = \lambda^{-1} f(\boldsymbol{w}^\star) - (m\lambda)^{-1}\left\langle \mathbf{1}, (\mathbf{1} - \boldsymbol{X}^\top \boldsymbol{w}^\star)_+ \right\rangle \\
&\leq \lambda^{-1} f(\boldsymbol{w}^\star) = \lambda^{-1} g(\boldsymbol{\alpha}^\star) \\
&= -\|\Pi_{\mathcal{S}}(\boldsymbol{v}(\boldsymbol{\alpha}^\star))\|^2/2 + (m\lambda)^{-1}\left\langle \boldsymbol{\alpha}^\star, \mathbf{1} \right\rangle \leq 1/(2\lambda)
\end{aligned}$$

を得る．ただし，最後の不等号を得るのに，$\boldsymbol{w}^\star = \Pi_{\mathcal{S}}(\boldsymbol{v}(\boldsymbol{\alpha}^\star))$ を用いた．よって，$\boldsymbol{w}^\star \in \mathcal{C}$ が成立する．あとは，文献 [2] と同様な論法で Theorem 1 を証明できる． □

図 1: 汎化性能の比較．

3. 大腸菌数予測への応用

河川における大腸菌数を水門水質データから予測する問題を考える．水門水質データとしては，WT, pH, EC, DO, SS, BOD, TN, TP, 流量を使用する．ただし，pH は $\mathrm{pH}_+ \leftarrow \max(0, \mathrm{pH} - 7)$ および $\mathrm{pH}\_ \leftarrow \max(0, 7 - \mathrm{pH})$ によって，2 つの説明変量 pH_+ および pH_ に分解した．水門水質データと大腸菌の関係から，WT, EC, SS, BOD, TN, TP は大きいほど，pH_+, pH_，DO，流量は小さいほど大腸菌が増えることが，すでに分かっている．このことから，SVM のパラメータ $\boldsymbol{w}$ の符号を次のように制限することとした：

- 説明変量 WT, EC, SS, BOD, TN, TP の係数 w_h は非負．
- 説明変量 pH_+, pH_, DO, 流量 の係数 w_h は非正．

河川における大腸菌数を 1 回測定するには 24 時間程度要し，試薬代は 6 千円かかる．本研究では，異なる場所と日時において 177 回大腸菌数を測定し，その測定場所と日時の水門水質データを収集した．このうち，10 個を訓練用データとして無作為に選び，符号制約 SVM (SR-SVM) と従来の SVM との比較を行った．残り 167 個を評価用データとして，PRBEP(Precision Recall Break Even Point) を算出した．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，Pegasos 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．これを河川大腸菌の予測問題に応用し，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] E. Hazan et al. Logarithmic regret algorithms for online convex optimization. *Mach Learn*, 69(2):169–192, 2007.

[2] S. Shalev-Shwartz et al. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, pages 807–814, New York, NY, USA, 2007. ACM.

[3] T. Kato et al. Estimation of concentration ratio of indicator to pathogen-related gene in environmental water based on left-censored data. *Journal of Water and Health*, 14(1):14–25, Feb 2016.