G-018

ヒト多能性幹細胞による化合物毒性予測アルゴリズムの検討
Study on compound toxicity prediction algorithm using human pluripotent stem cells

持田 陸斗 * 中島 未椰 * 加藤 毅 †
Rikuto Mochida Miya Nakajima Tsuyoshi Kato

1. はじめに

創薬は，あらゆる研究開発の中でも最も多くの時間と費用を要する分野の一つである．薬剤開発には，約 9〜17 年の期間と約 500 億円もの費用がかかり，臨床試験に到達した薬の 9 割は撤退を余儀なくされている．撤退の原因の 3 割は化合物の毒性である．撤退を未然に防ぐには，臨床試験の前に毒性を予測し，スクリーニングしておくことが重要である．従来，化合物の毒性予測は，主に動物実験によって，行われていた．しかし，近年，倫理的な問題から多くの国で規制や禁止が計画，もしくはすでに実施されている．その他にも，ヒトと動物との作用機序の違いにより，動物実験による毒性予測を行って承認された化合物が，ヒトに対して毒性を示す可能性がある．

動物実験の代替として，ヒト多能性幹細胞を使った毒性予測が注目を集めている．すなわち，**iPS 細胞**や**ES 細胞**などの多能性幹細胞に化合物を暴露させ，それによる遺伝子発現量の変化から，機械学習モデルを介して毒性を予測する．Yamane ら [2] は遺伝子発現の変化から遺伝子ネットワークを構成して予測するというアプローチを示し，ヒト多能性幹細胞の有用性を明らかにした．

多能性幹細胞を用いた方法はデータを大規模化が容易ではないという短所を抱えている．現在，発現データを得るための RNAseq は約４万円以上費用がかかる．化合物ごと，また経過時間ごとに遺伝子発現量を測定し，リピート実験も実施するとなると，膨大な費用と時間を費やすことになる．遺伝子発現量を得るためのプロトコルも完全には標準化されていないため，異なる研究組織で測定した発現データを安易に訓練用データに統合することはできない．データの大規模化に障壁があるため，比較的少量のデータでも汎化する線形予測器やカーネル法を用いたとしても，機械学習に十分なデータ規模に至りづらい．

本論文では，深層学習で成功している **Mixup データ拡張** [3] をカーネル予測器に適用することで予測性能を向上できることを示す．Mixup は，訓練用例題をクラスラベルも含めて凸結合を取ることでデータを水増しする方法である．本研究では，予測器の学習のために，**確率的双対座標上昇法** (**SDCA**) [1] を適用した．この研究を通して，以下の成果を得た．

- Mixup 法適用によって生じる**中間値クラスラベル**から**再生核ヒルベルト空間**上で学習する **Exact SDCA** を開発した．筆者らの知る限り，これまでは，中間値クラスラベルに対するカーネル法の学習アルゴリズムは十分解析されてこなかった．SDCA の適用により，最適化のためのハイパーパラメータが不要になる．これは対話的な最適化を要する **SGD**(確率的勾配降下法) よりも利便性が高い．
- 従来，Exact SDCA を適用するには，各反復で解くべき部分問題の目的関数が放物線になる必要があった．一般に，中間値クラスラベルを用いると部分問題の目的関数は放物線にならない．しかし，筆者らは，ある条件下では中間値クラスラベルを用いない場合と同じ計算コストで Exact SDCA の各反復を実行できることを発見した．
- 本研究で開発した Exact SDCA は，**近似 SDCA** や SGD よりも顕著に少ないエポック数で最適解に到達できることを数値実験により確かめた．
- 多能性幹細胞を用いた毒性予測において，遺伝子ネットの多重化と Mixup の適用が有効であることを実データを用いた実験によって明らかにした．

2.Mixup データを用いた学習問題

Mixup 法は，訓練用データセットから無作為に選択した 2 例題 $(\boldsymbol{x}_i^{\mathrm{o}}, y_i^{\mathrm{o}}), (\boldsymbol{x}_j^{\mathrm{o}}, y_j^{\mathrm{o}}) \in \mathbb{R}^d \times \{\pm 1\}$，およびベータ分布から生成した $\eta \in [0,1]$ を使って，新しい例題

$$\boldsymbol{x}_{\mathrm{new}} = (1-\eta)\boldsymbol{x}_i^{\mathrm{o}} + \eta \boldsymbol{x}_j^{\mathrm{o}}, \quad y_{\mathrm{new}} = (1-\eta)y_i^{\mathrm{o}} + \eta y_j^{\mathrm{o}}$$

を生成する．これを繰り返すことでデータの水増しを実現する．すると，新たに生成した化合物のクラスラベル y_{new} は 2 値ではなく，中間値になっていることに注意されたい．

Mixup データ拡張によって得たデータセット $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ から予測器 $f : \mathbb{R}^d \to \mathbb{R}$ を訓練したい．Mixup データから予測器 f を学習するには，次式で定義される Mixup 損失関数 $\phi_{\mathrm{mup}}(\cdot\,; y) : \mathbb{R} \to \mathbb{R}$ を用いる：

$$\phi_{\mathrm{mup}}(s\,; y) := \frac{1+y}{2}\phi(s) + \frac{1-y}{2}\phi(-s).$$

ここで用いられている $\phi : \mathbb{R} \to \mathbb{R}$ は中間値クラスラベルを用いない損失関数である．たとえば，ロジスティック回帰を実現する 2 値交差エントロピー損失 $\phi_{\mathrm{bce}}(s) := \log(1+\exp(-s))$ や，SVM を実現する平滑ヒンジ損失

$$\phi_{\mathrm{smh}}(s) := \begin{cases} -x + 1 - \frac{\gamma_{\mathrm{sm}}}{2} & \text{if } x < 1 - \gamma_{\mathrm{sm}}, \\ \frac{1}{2\gamma_{\mathrm{sm}}}(x-1)^2 & \text{if } 1 - \gamma_{\mathrm{sm}} \le x < 1, \\ 0 & \text{if } 1 \le x. \end{cases}$$

*群馬大学理工学府
†群馬大学情報学部

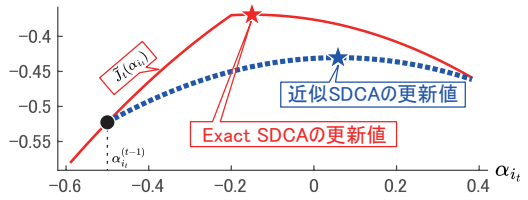

図 1: Exact SDCA と近似 SDCA. Exact SDCA は $\widetilde{J}_t(\alpha_{i_t})$ を最大化するよう双対変数を更新する．近似 SDCA は $\widetilde{J}_t(\alpha_{i_t})$ を近似する放物線を最大化する．

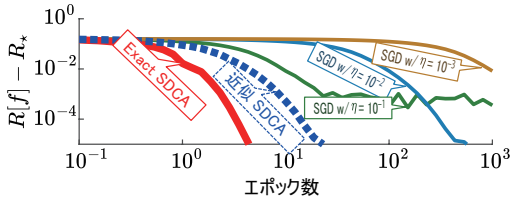

図 2: 収束速度の比較．η は SGD のステップサイズ．

などが用いられる．ただし，$\gamma_{\mathrm{sm}} \in (0,1)$ は定数である．カーネル法では，**正定値カーネル関数** $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (e.g., RBF カーネル，多項式カーネル) から構成される再生核ヒルベルト空間 $\mathcal{H}_\kappa$ の中から正則化経験リスク $R[f]$ を最小にする予測器 f を探すことになる．すなわち以下の最小化問題を解く：

$$\begin{aligned} &\min \quad R[f] \qquad \text{wrt} \quad f \in \mathcal{H}_\kappa \\ &\text{where} \quad R[f] := \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{n}\sum_{i=1}^{n} \phi_{\mathrm{mup}}(f(\boldsymbol{x}_i)\,;\, y_i) \end{aligned}$$

ただし，$\lambda > 0$ は正則化定数である．

3. 最適化アルゴリズム

本研究では，正則化経験リスク $R[f]$ の最小化のために SDCA を適用する．SDCA は $R[f]$ を直接最小化する代わりに，その**フェンシェル双対** $D : \mathbb{R}^n \to \mathbb{R}$ を最大化する：

$$D(\boldsymbol{\alpha}) := -\frac{\lambda}{2}\|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_\kappa}^2 - \frac{1}{n}\sum_{i=1}^{n} \phi^*_{\mathrm{mup}}(-\alpha_i) \qquad (1)$$

ただし，$\phi^*_{\mathrm{mup}} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ は ϕ_{mup} の凸共役を表す．また，$f_{\boldsymbol{\alpha}} := \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x}_i, \cdot) \in \mathcal{H}_\kappa$ とおいた．$\boldsymbol{\alpha}_\star \in \mathbb{R}^n$ を $D(\boldsymbol{\alpha})$ を最大化する解とすると，$R[f]$ を最小化する解は $f_\star := f_{\boldsymbol{\alpha}_\star}$ と表される．Exact SDCA は，各反復 t で無作為に例題 $i_t \in [n]$ を選び，$\widetilde{J}_t(\alpha_{i_t}) := D(\boldsymbol{\alpha}^{(t-1)} + (\alpha_{i_t} - \alpha_{i_t}^{(t-1)})\boldsymbol{e}_{i_t})$ を最大化するように α_{i_t} を更新する．ただし，$\boldsymbol{\alpha}^{(t-1)} \in \mathbb{R}^n$ は第 $(t-1)$ 反復における双対変数の値である．ϕ が平滑ならば ϕ_{mup} も平滑になることが容易に確かめられ，よって最適解への線形収束が保証される [1]．各反復の計算コストは $\widetilde{J}_t(\alpha_{i_t})$ の最大化問題をいかに効率的に解けるかに依存する．本研究では，次の定理を発見した：

表 1: F1 スコアによる毒性予測性能の比較.

手法	神経毒性	肝毒性
従来法 [2]	0.588	0.703
ネット多重化 (Mixup なし)	0.800	0.897
ネット多重化 (Mixup あり)	**0.838**	**0.906**

定理 1. ϕ に平滑ヒンジを用いる場合，関数 $\widetilde{J}_t(\alpha_{i_t})$ は，**凹区分２次関数**で表される．その区間数は高々 2 である．よって，各反復の計算コストを $O(n)$ に抑えたまま *Exact SDCA* の更新を実現できる．

Proof sketch: ϕ が平滑ヒンジの時，凸共役 ϕ^*_{mup} は凹区分２次関数になる．よって，$\widetilde{J}_t(\alpha_{i_t})$ も凹区分２次関数になる．その係数は $O(n)$ の計算コストで求まることも導くことができる．凹区分２次関数の最適化は $O(1)$ で済む． □

従来，$\widetilde{J}_t(\alpha_{i_t})$ が単一の放物線ではないとき，近似 SDCA が用いられていた．図 1 に示すように，近似 SDCA では，最適ではない位置に更新されてしまう．定理 1 は，$\widetilde{J}_t(\alpha_{i_t})$ が放物線ではないのに，放物線の場合と同じ計算コストで Exact な更新ができることを示したことになる．

4. 実験

収束実験：数値実験によって，Exact SDCA，近似 SDCA, SGD を比較した．例題数を $n = 1{,}000$ とし，正則化定数には $\lambda = 1/n$ を用いた．図 2 に横軸エポック数，縦軸目的ギャップ $R[f] - R_\star$ として学習の様子をプロットした．ただし，$R_\star := \min_f R[f]$ である．Exact SDCA は $R[f] - R_\star < 10^{-5}$ に達するのに 5.34 エポックで済んだが，近似 SDCA は 43.29 エポックも要した．SGD は $\eta = 0.01$ を選んだ場合だけ 1,000 エポック以内に $R[f] - R_\star < 10^{-5}$ に達した．

毒性予測実験：従来法である StemPanTox 法 [2] では，単一の遺伝子ネットしか使っていなかった．これを多重化した方法，さらに Mixup で 100 例題追加した方法を比較した．[2] で用いられていた 24 化合物の神経毒性と肝毒性を一つ抜き法で予測し，F1 スコアで評価した．その結果を表 1 に示す．遺伝子ネットを多重化して，Mixup でデータ拡張することにより，予測性能が向上することが確かめられた．

謝辞: 本研究は JSPS 科学研究費 22K04372 および 23H03555 の助成を受けた．

参考文献

[1] S. Shalev-Shwartz & T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *JMLR*, 14(1):567–599, 2013.

[2] J. Yamane et al., StemPanTox: A fast and wide-target drug assessment system for tailor-made safety evaluations using personalized iPS cells. *iScience*, 25(7):104538, 2022.

[3] H. Zhang et al, mixup: Beyond empirical risk minimization. In *ICLR*, 2018.