

半教師あり学習を用いた医療データにおける標的遺伝子の探索

Semi-supervised learning based target gene selection for medical data

関 爽人^{*1}
Akito Seki

秋田 大空^{*1}
Hirotaka Akita

鹿島 久嗣^{*1*2}
Hisashi Kashima

山田 誠^{*1*2}
Makoto Yamada

^{*1}京都大学
Kyoto University

^{*2}RIKEN AIP

A therapeutic method that controls disease progression by administering a drug which inhibits the function of a particular molecule is called molecularly targeted therapy. Since the number of genes which can be targeted in this method is quite large, it is impossible to verify all of them by human hands. Therefore, in this study, we see the task of finding new target genes as a problem of machine learning and propose a framework in order to solve this problem. In the experiment, we used three datasets and examined the proposed method. As a result, the method outperformed the baseline even when we used any dataset.

1. はじめに

ある特定の分子の働きを阻害することにより病気の進行を抑制する治療法は分子標的治療と呼ばれ、主にがん治療の分野で用いられている。医療データを用いて分子標的治療における新たな標的遺伝子を発見することは、その病気に対する新薬の開発に必要となる、極めて重要なタスクである。ただし、標的となり得る遺伝子の候補を人間の手で網羅的に検証するには、多くのコストと時間を要する。そのため新たな標的を発見するためにはまず、遺伝子の発現量に関するデータを用いて、治療において重要だと考えられる遺伝子を特定する必要がある。このような課題は、機械学習やデータマイニングの分野で長年研究されてきた、特徴選択の問題であると考えられることができる。しかし、一般的に特徴選択を用いた手法は人間の知識を利用できないため、治療に重要な標的遺伝子を発見できない恐れがある。そこで、特徴選択で選んだ少数の遺伝子をケミカルスクリーニングなどによって人手で評価し、得られたデータを解析することで新たな標的遺伝子を発見することを考える。

本研究が提案するフレームワークについて以下に述べる。はじめに、遺伝子の発現量に関するデータセットに対して Hilbert-Schmidt Independence Criterion (以下、HSIC) Lasso [Yamada 14] を用いて特徴選択を行い、疾患の有無と強い関係を持っている遺伝子を選択する。この手法は医療データのような複雑なデータに対しても適用できる点、冗長な特徴量を排除できる点において優れている。次に、特徴選択の結果選ばれた遺伝子を、専門家がケミカルスクリーニングなどを用いて分析する。このステップにより、遺伝子の発現量データと、各遺伝子に対応するラベルから構成されるデータセットを得ることができる。最後に、このデータセットを用いて半教師あり学習を行う。本研究では、Virtual Adversarial Training (以下、VAT) [Miyato 18] という手法を用いる。これは、独自の正則化項を導入することによりネットワークの汎化性能を向上させる手法であり、半教師あり学習において優れた結果を示すことが知られている。学習を行ったのち、新たな標的遺伝子の予測を行う。具体的には、ラベル付けされていない各遺伝子に対して重要度を表すスコアを算出し、そのスコアが特に大きいものを新たな標的遺伝子の候補として扱うものとする。この予測の精度を評価するためには、候補となった遺伝子に対して再び専門家による評価を行い、各遺伝子が実際に標的遺伝子と

なるか否かを検証する必要がある。

本研究の主な貢献としては、専門家による評価を用いた、新たな標的遺伝子発見のためのフレームワークの構築を行ったこと、また、上記のフレームワークに基づいた半教師あり学習を提案し、実験によってその有効性について検証したことが挙げられる。また、VAT などの手法による半教師あり学習は一般的に分類問題に適用されており、我々の知る限りでは、半教師あり学習を用いて新たな標的遺伝子を探索する取り組みは初めてである。

本研究における実験では、ウェブ上に公開されている3つのデータセットを用いて、新たな標的遺伝子の予測精度について検証を行った。この実験においては、特徴選択によって得られた特徴量がすべて真の標的遺伝子であることを仮定し、これらの特徴量を半教師あり学習を用いて識別できるかを評価した。その結果、いずれのデータセットを用いた場合も、訓練データの数の大小に関わらず、提案手法がベースラインの性能を上回ることが示された。

2. 関連研究

Least absolute shrinkage and selection operator (以下、Lasso) [Tibshirani 96] は、L1 正則化を用いた回帰の代表的な手法である。L1 正則化は、モデルのパラメータを推定する際に使用されることで過学習を防ぎ、モデルの汎化性能を向上させる効果が期待できる。Lasso はスパース推定を行うため、特徴選択のアルゴリズムとしても一般的に使用されている。ただしこの手法には、特徴量の数を d 、データ数を n で表すとき、 $d > n$ ならば、高々 n 個の特徴量しか選択できないという欠点があることが知られている。さらに、Lasso は線形手法であり、医療データのような複雑なデータでは重要な特徴量を選択できないという問題がある。

HSIC Lasso [Yamada 14] は、冗長性を排除した特徴選択を行える上に、大域的最適解を得ることができる手法である。また、この手法は非線形な特徴選択を行うため、複雑なデータを用いた際も重要な特徴量を選択できると考えられる。一方、HSIC Lasso によって特徴選択を行うだけでは人間の知識を利用できないため、専門家による評価も併用して行うことが必要となる。

半教師あり学習は、一般に少数の教師ありデータと多数の

教師なしデータから構成されるデータセットを用いて行われる学習を指す。このようなデータセットが多く存在する画像処理や情報検索、生命情報科学の分野において、半教師あり学習は広く適用されている [Chapelle 09]。

半教師あり学習において優れた結果を示す手法として、Miyato らによって考案された VAT [Miyato 18] がある。この手法は、教師ありデータと教師なしデータの両方を用いて計算される独自の正則化項を目的関数に導入することにより、ネットワークの汎化能力を向上させている。

3. 問題設定

まず、問題設定における表記法について述べる。本研究においては医療データ、特に遺伝子の発現量に関するデータを取り扱うことを想定している。この時、 $\mathbf{x} = [x^{(1)}, x^{(2)}, \dots, x^{(d)}]^T \in \mathbb{R}^d$ を一人の被験者における各遺伝子の発現量を表すベクトル、 $y \in \mathbb{R}$ を疾患の有無など被験者の状態を表すラベルであるとする。任意の行列 $\mathbf{A}$ に対して、 $\mathbf{A}^T$ は $\mathbf{A}$ の転置行列を表しているものとする。いま、互いに独立で同一の分布に従う n 組のデータが与えられていると仮定し、これを $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \dots, n\}$ と表す。上記の $\mathbf{x}$ を用いて、遺伝子データ全体を $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ と表す。また、上記の y を用いて、被験者のラベル全体を $\mathbf{y} = [y_1, y_2, \dots, y_n]^T \in \mathbb{R}^n$ と表す。

本研究では、このような遺伝子データと被験者のラベルに加え、各特徴量に付与されたラベルについても情報を得ることができると仮定する。各ラベルを $t_i \in \{-1, 0, 1\}$ として定義し、 t_i は専門家が行った分析の結果得られるラベルであるとする。値が 1 の場合はその特徴量が有効であることを表し、-1 の場合は有効でないことを表し、0 は有効であるかが不明であることを表している。また、 $\mathbf{u} = [u^{(1)}, u^{(2)}, \dots, u^{(n)}]^T \in \mathbb{R}^n$ を各特徴量データとして定義すると、特徴量データとその特徴量に付与されたラベルの組 $\{(\mathbf{u}_j, t_j) | \mathbf{u}_j \in \mathcal{U}, t_j \in \mathcal{T}, j = 1, \dots, d\}$ を得ることができる。具体的には、 $\mathbf{u}$ は一つの遺伝子の各被験者における発現量を表すベクトルであり、 t は遺伝子が標的であるか否かを表すラベルである。ただしラベル t については、専門家による分析を行った後のみ利用することができるものとする。上記の t を用いて、各特徴量に対応するラベル全体を $\mathbf{t} = [t_1, t_2, \dots, t_d]^T \in \mathbb{R}^d$ と表す。また、上記の $\mathbf{u}$ を用いて、特徴量データ全体を $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d] \in \mathbb{R}^{n \times d}$ と表すと、 $\mathbf{X}^T = \mathbf{U}$ が成立する。

本研究における入出力を以下に示す。

- 入力：遺伝子データ $\mathbf{X}$ 、被験者のラベル $\mathbf{y}$ 、特徴量データ $\mathbf{U}$ 、特徴量のラベル $\mathbf{t}$
- 出力：有効であるか不明な遺伝子 $\mathbf{u}_j$ に対する、重要度を表すスコア $\hat{t}_j \in [0, 1]$

出力のスコアについては、その値が 1 に近いほど、その特徴量がより重要であると予測したことを表している。

4. 提案手法

本研究では、医療データにおける新たな標的遺伝子の発見を目的とした新たなフレームワークを提案する。これは大きく分けて、遺伝子データを用いた特徴選択、選択された特徴量に対する専門家による評価、半教師あり学習による新たな標的遺伝子の予測という 3 つのステップから構成される。これらについて、以下に詳細を述べる。

4.1 遺伝子データを用いた特徴選択

このステップにおける入力、各被験者の遺伝子の発現量を記録したデータと、疾患の有無など各被験者の状態を表すラベルであり、それらの組を $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \dots, n\}$ と表す。このような入力を用いて、その疾患と強い関係がある特徴量を選択することをタスクとする。ここで、第 2 ステップの専門家による評価においては、検査できる遺伝子の数が限られているため、不要で冗長な特徴量を可能な限り排除した選択を行うことが求められる。また、遺伝子の発現量データは複雑であるという特徴がある。これらの条件を考慮し、非線形な特徴選択を行う最先端のアルゴリズムとして知られる HSIC Lasso [Yamada 14] を用いる。

まず、HSIC Lasso における最適化問題を以下に示す。

$$\min_{\alpha \in \mathbb{R}^d} \|\tilde{\mathbf{L}} - \sum_{k=1}^d \alpha_k \tilde{\mathbf{K}}^{(k)}\|_F^2 + \lambda \|\alpha\|_1, \quad s.t. \quad \alpha_1, \dots, \alpha_d \geq 0 \quad (1)$$

$\lambda \geq 0$ は正則化項の係数を表すパラメータである。 $\|\cdot\|_1$ は L1 ノルム、 $\|\cdot\|_F$ はフロベニウスノルム ($\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^T)}$) を表し、 $\tilde{\mathbf{K}}$ と $\tilde{\mathbf{L}}$ はそれぞれ、以下の式で与えられる行列を表している。

$$\tilde{\mathbf{K}} = \frac{\mathbf{H}\mathbf{K}\mathbf{H}}{\|\mathbf{H}\mathbf{K}\mathbf{H}\|_F}, \quad \tilde{\mathbf{L}} = \frac{\mathbf{H}\mathbf{L}\mathbf{H}}{\|\mathbf{H}\mathbf{L}\mathbf{H}\|_F} \quad (2)$$

$\mathbf{K}$ と $\mathbf{L}$ はそれぞれ、遺伝子データとラベルに関するグラム行列を表し、 $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ は中心化行列である。 $\mathbf{I}$ は単位行列であり、 $\mathbf{1}$ は全ての要素が 1 である正方行列を表している。式 (2) 及び $\mathbf{H}$ の性質より、 $\tilde{\mathbf{K}}$ と $\tilde{\mathbf{L}}$ は以下のような性質を持つ。

$$\mathbf{1}_n^T \tilde{\mathbf{K}} \mathbf{1}_n = \mathbf{1}_n^T \tilde{\mathbf{L}} \mathbf{1}_n = 0 \quad (3)$$

$$\|\tilde{\mathbf{K}}\|_F^2 = \|\tilde{\mathbf{L}}\|_F^2 = 1 \quad (4)$$

4.2 選択された特徴量に対する専門家の評価

まず、第 1 ステップで選択された特徴量に対する評価を専門家に依頼する。具体的には、各遺伝子に対して、それが標的として有効であると判断すればラベルの値を 1 に設定し、逆に有効ではないと判断すればラベルを -1 に設定する。このような検査には多くのコストと時間を要するため、検査できる遺伝子の数が限られているという制約がある。

4.3 半教師あり学習による新たな標的遺伝子の予測

第 2 ステップの結果得られたデータ $\{(\mathbf{u}_j, t_j)_{j=1}^d\}$ を入力とする。 $\mathbf{u}_j$ は遺伝子の各被験者における発現量を表す。 $t_j \in \{-1, 0, 1\}$ は各遺伝子に対応するラベルであり、評価を行わなかった遺伝子には 0 が付与されている。このステップでは、半教師あり学習において優れた結果を示す VAT [Miyato 18] を用いる。VAT は、独自の正則化項を導入することによりネットワークの汎化性能が向上させる手法であり、パラメータの更新を繰り返して学習を行う。この手法について以下に述べる。

まず、Local Distributional Smoothness (以下、LDS) という関数を、入力データ $\mathbf{u}$ とパラメータ θ について、以下のよう

$$LDS(\mathbf{u}, \theta) := D[p(t|\mathbf{u}, \hat{\theta}), p(t|\mathbf{u} + r_{\text{adv}}, \theta)] \quad (5)$$

t は出力ラベル、 $\hat{\theta}$ は現在のパラメータの値を表し、 $p(t|\mathbf{u}, \theta)$ は事後分布を表す。 $D[p, p']$ は、2 つの分布 p と p' の差異を計る尺度を表しているが、本研究では D としてクロスエントロ

ピーを用いる。また、式 (5) で用いられている r_{vadv} は、以下の形で定義される。

$$r_{vadv} := \arg \max_{r: \|r\|_2 \leq \epsilon} D[p(t|\mathbf{u}, \hat{\theta}), p(t|\mathbf{u} + r, \hat{\theta})] \quad (6)$$

ϵ はハイパーパラメータであり、 $\|\cdot\|_2$ は L2 ノルムを表している。この設定のもとで、 r_{vadv} は、2 つの分布 $p(t|\mathbf{u}, \hat{\theta})$ と $p(t|\mathbf{u} + r, \hat{\theta})$ の差異がもっとも大きくなる時の摂動 r を表し、LDS は、入力データ $\mathbf{u}$ 周辺での事後分布の滑らかさを表している。

次に、式 (5) で定義された LDS を用いて正則化項 R_{vadv} を定義する。

$$R_{vadv}(\theta) := \frac{1}{N_l + N_{ul}} \sum_{\mathbf{u} \in \mathcal{U}_l, \mathcal{U}_{ul}} LDS(\mathbf{u}, \theta) \quad (7)$$

ここで N_l は、専門家による評価の結果ラベル付けされたデータ、つまり、ラベルの値が 1 もしくは -1 であるデータの数を表している。一方 N_{ul} は、専門家による評価を行わなかったデータ、つまり、ラベルの値が 0 であるデータの数を表している。最後に、式 (7) で定義された R_{vadv} を用いて、目的関数全体を以下のように表す。

$$\frac{1}{N_l} \sum_{(t, \mathbf{u}) \in \mathcal{T} \times \mathcal{U}_l} \ell(t, f(\mathbf{u}, \theta)) + \alpha R_{vadv}(\theta) \quad (8)$$

$f(\mathbf{u}, \theta)$ は、 θ をパラメータとして用いるモデルを表しており、 α はハイパーパラメータである。 $\ell(t, f(\mathbf{u}, \theta))$ は、ラベルありデータについての負の対数尤度を表す。ここで、式 (8) を最小化するように学習を行うことを考えると、この問題は誤差逆伝播 [Rumelhart 86] によって効率的に最適化を行うことができる。学習が終了したのち、モデル $f(\mathbf{u}, \theta)$ を用いて新たな標的遺伝子の予測を行う。この際、専門家による評価を行わなかった各遺伝子に対して、重要度を表すスコアを算出する。スコアの値は 0 以上 1 以下であるとし、その値が 1 に近いほどその遺伝子がより重要であると予測したことを表している。最後に、これらの遺伝子のうち、スコアが特に大きいものを新たな標的遺伝子として考える。

5. 評価実験

5.1 実験概要

本研究が提案するフレームワークでは、遺伝子の発現量に関するデータを用いて特徴選択を行い、選択された遺伝子に対する専門家の評価によって、それらの遺伝子にラベル付けを行っている。しかし本研究の実験では、専門家による評価を実際に行うことはできなかったため、以下のような方法を用いている。

まず、特徴選択の結果選ばれたすべての遺伝子に対してラベルの値を 1 に設定する。そして、特徴選択の際選ばれなかった遺伝子に対してはラベルの値を -1 に設定する。したがって、各遺伝子に対して与えられるラベルは疑似ラベルとなる。このように、すべての遺伝子に対するラベル付けを行った上で、ラベルの値が 1 であるデータとラベルの値が -1 であるデータをそれぞれ同数選択し、これらを訓練データとして用いる。一方、訓練データとして選択されなかったデータは全て、テストデータとして用いる。ただしモデルの訓練の際、テストデータのラベル情報は利用できないが、これらを教師なしデータとして利用することはできるものとする。

5.1.1 実験設定

はじめに、入力として与えられるデータセットに対して特徴選択を行うが、第 3 章で述べたように HSIC Lasso [Yamada 14] を用いる。その後、先に述べた方法で各遺伝子にラベルを付与し、訓練データとテストデータに分割する。次に、訓練データとラベル情報を取り除いたテストデータを用いてモデルの学習を行う。最後に、テストデータを用いてモデルの性能を測定する。訓練データとテストデータの分割については偏りが生じる恐れがあるため、データの分割を変えながらテストを 10 回行うものとする。

半教師あり学習におけるモデルとして、4 層のニューラルネットワークを用いる。提案手法とベースラインにおいては共通したネットワークを用いるものとする。ニューラルネットワークの入力層のユニット数は、各実験における被験者の数と等しいものとし、2 つある隠れ層のユニット数は 100、出力層のユニット数は 2 とする。出力層における活性化関数としては softmax 関数を用い、その他の層における活性化関数としては ReLU [Glorot 11] を用いた。また、ネットワークの最適化手法として確率的勾配降下法を使用した。

5.1.2 ベースライン

本項では、ベースラインとして用いる手法について述べる。入力として与えられるデータセットに対する特徴選択や、各遺伝子へのラベル付けの方法などはすべて、提案手法と同様に行われるものとする。前節でも述べたように、新たな標的遺伝子を予測するための学習の際に、モデルとして用いるニューラルネットワークについても同様である。提案手法と異なるのは、学習を行う際に用いられる目的関数である。その関数は以下の形で与えられる。

$$\frac{1}{N_l} \sum_{(t, \mathbf{u}) \in \mathcal{T} \times \mathcal{U}_l} \ell(t, f(\mathbf{u}, \theta)) \quad (9)$$

5.1.3 評価指標

本実験では正例と負例の数に偏りがあるデータが使用されるため、評価指標は ROC 曲線の AUC を用いる。

5.2 データセット

本研究が提案するフレームワークの有効性を検証するため、3 つのデータセットを用いて実験を行う。実験において用いる 3 つのデータセットは全て、FEATURE SELECTION DATASETS^{*1} に公開されているものである。これらについて簡単に述べる。leukemia データセットは、各被験者の遺伝子の発現量を 3 値化したデータと、各被験者が患者であるか否かを表す二値のラベルからなる。発現量は -2, 0, 2 のいずれかとして与えられている。72 人の被験者のうち、患者である 25 人を正例として扱い、残りの 47 人を負例として扱う。発現量のデータが存在する遺伝子の数は 7,070 個である。Prostate_GE データセットは、各被験者の遺伝子の発現量を表したデータと、各被験者が患者であるか否かを表す二値のラベルからなる。102 人の被験者のうち、患者である 50 人を正例として扱い、残りの 52 人を負例として扱う。発現量のデータが存在する遺伝子の数は 5,966 個である。TOX_171 データセットは、各被験者の遺伝子の発現量を表したデータと、各被験者の状態を表す 4 値のラベルからなる。各クラスを構成する被験者の数はそれぞれ、39 人、42 人、45 人、45 人であり、その総数は 171 人である。発現量のデータが存在する遺伝子の数は 5,748 個である。

*1 <http://featureselection.asu.edu/datasets.php#>

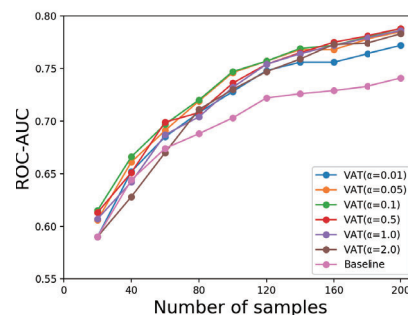
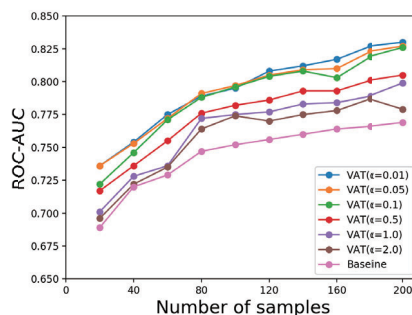
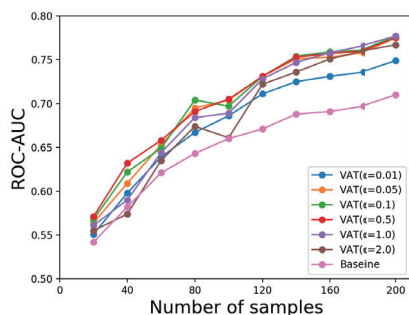


図 1: leukemia データを用いた実験の結果

図 2: Prostate_GE を用いた実験の結果

図 3: TOX_171 を用いた実験の結果

5.3 実験結果

まず、leukemia データセットを用いた実験の結果について述べる。特徴選択を行った結果、7,070 個の特徴量のうち 604 個が選択され、それらのラベルの値は 1 に設定された。このデータを用いて提案手法とベースラインの性能を評価したものを図 1 に示す。次に、Prostate_GE データセットを用いた実験の結果について述べる。特徴選択を行った結果、5,966 個の特徴量のうち 398 個が選択され、それらのラベルの値は 1 に設定された。このデータを用いて提案手法とベースラインの性能を評価したものを図 2 に示す。最後に、TOX_171 データセットを用いた実験の結果について述べる。特徴選択を行った結果、5,748 個の特徴量のうち 564 個が選択され、それらのラベルの値は 1 に設定された。このデータを用いて提案手法とベースラインの性能を評価したものを図 3 に示す。

以上のように、いずれのデータセットを用いた場合も、提案手法はベースラインを上回る結果を示した。特に Prostate_GE データセットを用いた実験においては、提案手法がベースラインを下回ることが一度もない点、また、他のデータセットを用いた場合と比較して提案手法の ROC 曲線の AUC が一貫して高い値を取る点において、提案手法の有効性が示された。一方で、用いたデータセットによって提案手法が高い性能を示す時のパラメータ α の値が異なったため、各データセットに適した値をチューニングによって検証する必要があると感じた。

6. むすび

本研究では、ある疾患における新たな標的遺伝子を発見することを目的とした新たなフレームワークを提案した。実験においては、3 つの遺伝子の発現量に関するデータセットを用いて提案手法の性能をベースラインと比較し、提案手法の有効性について検証した。

一方、本研究の課題としては、実験において専門家による評価を行えなかったこと、またこれに関連して、実際に専門家による評価を運用した際の提案手法の有効性を示せなかったことが挙げられる。今後は、専門家による協力のもと提案手法を用いた実験を行い、新たな標的遺伝子を発見することを目標とする。また今回の実験では、ハイパーパラメータのチューニングをせず、目的関数における正則化項の係数 α のみ値を変えながら実験を行った。本来は教師ありデータを訓練データ、検証データ、テストデータの 3 つに分割して実験を行うべきであるが、そのためには、少数の教師ありデータを有効に活用するための工夫が必要となるため、これを今後の課題としたい。

また、本研究が提案したフレームワークを拡張として、フレームワークの第 2 ステップと第 3 ステップを交互に繰り返

し、新たな標的遺伝子の探索を何度も行うというものが考えられる。このような拡張を行うと、半教師あり学習によって重要度が高いと考えられる遺伝子を得たのち、再度専門家による評価を行えるので、これらの遺伝子を新たな教師ありデータとして用いることができる。したがって、標的遺伝子の探索の精度を向上させることができると考えられる。一方で、専門家による評価を繰り返すことで実験にかかる費用が増大するため、反復回数をできる限り少なく抑えることが重要となる。

参考文献

- [Chapelle 09] Chapelle, O., Scholkopf, B., and Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews], *IEEE Transactions on Neural Networks*, Vol. 20, No. 3, pp. 542–542 (2009)
- [Glorot 11] Glorot, X., Bordes, A., and Bengio, Y.: Deep sparse rectifier neural networks, *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323 (2011)
- [Miyato 18] Miyato, T., Maeda, S., Ishii, S., and Koyama, M.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning, *IEEE transactions on pattern analysis and machine intelligence* (2018)
- [Rumelhart 86] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, *nature*, Vol. 323, No. 6088, p. 533 (1986)
- [Tibshirani 96] Tibshirani, R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288 (1996)
- [Yamada 14] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M.: High-dimensional feature selection by feature-wise kernelized lasso, *Neural computation*, Vol. 26, No. 1, pp. 185–207 (2014)