4J1-1in2

部分サンプリングに基づく特徴選択を用いたウイルス感染の予測法

Virus Infection Prediction with Feature Selection Based on Sub-sampling

佐藤 浩基 *1　志賀 元紀 *2*3
Hiroki Sato　Motoki Shiga

*1岐阜大学大学院自然科学技術研究科
Graduate School of Natural Science and Technology, Gifu University

*2岐阜大学工学部
Faculty of Engineering, Gifu University

*3科学技術振興機構さきがけ
PRESTO, JST

To prevent pandemics of respiratory viruses, early detection of individuals infected with viruses is an important task. Against this task, molecular measurements in a cell such as gene expression measurements are useful to detect minor changes caused by a virus infection. There exist several machine learning methods for this task but it is essentially difficult because of the small-N-large-P setting. We developed a new stability feature selection method with auxiliary group information based on a group sparse norm and random ensemble model. Our developed method is evaluated using both synthetic and real datasets.

1. はじめに

インフルエンザなどの呼吸器系ウイルスは感染能力が高いために，感染の早期発見が重要である．これに対して，採取した細胞の遺伝子発現量から，感染による体内活動の変化を早期に発見する試みがある [Zaas2009, Woods2013]．しかしながら，感染直後の体内環境の変化は微小であり，さらに，候補の特徴量となる遺伝子が多数であるものの多数の被験者に対する実験および観測を行えず困難である．選択された特徴量（遺伝子）を実験分野にフィードバックして解釈することが必要となるが，高次元における特徴選択を少ない標本数から行う状況においては，1 つの観測標本が追加されたり観測値が微小変動するだけでも，選択される特徴量が大きく変わってしまい，誤った結論を導く危険性が高まる．本研究では，スパース回帰における特徴選択の安定化法 [Nicolai 2010] を拡張し，標本の部分サンプリングおよび補助グループ情報をを用いた特徴選択の安定化法を提案する．人工データおよび呼吸器系ウィルスに関する実データを用いた数値実験により提案法を検証する．

2. 部分サンプリングおよびグループ補助情報を用いた安定な特徴選択

各被験者が P 個の遺伝子（特徴量）の発現量 $\boldsymbol{x} \in \mathbb{R}^P$ と呼吸器系ウィルスの感染を示す 2 値目的変数 $y \in \{0,1\}$ として，N 人の被験者の観測標本 $(\boldsymbol{x}_n, y_n)$, $n = 1 \ldots, N$ が与えられたとする．パラメータ $\boldsymbol{w}$ を用いれば，説明変数と目的変数の関係がロジット関数

$$f(\boldsymbol{x}) = \log \frac{q(y=1|\boldsymbol{x})}{q(y=0|\boldsymbol{x})} = \boldsymbol{x}^T \boldsymbol{w} \tag{1}$$

において線形モデルで仮定される．このとき，目的変数は

$$\hat{y}_n = q(y=1|\boldsymbol{x}_n) = \sigma(\boldsymbol{x}_n^T \boldsymbol{w}) \tag{2}$$

となる．ただし，シグモイド関数 $\sigma(a) = \frac{1}{1+\exp(-a)}$ である．パラメータ $\boldsymbol{w}$ は損失関数（尤度関数の負の対数）

$$L(\boldsymbol{w}) = -\sum_{n=1}^{N} \log q(y_n|\boldsymbol{x}_n, \boldsymbol{w}) \tag{3}$$

の最小化に基づき決定される．

ところで，各遺伝子の機能などの情報を用いれば，特徴量をグループ化できる．各遺伝子は複数の機能グループに属するため，与えられるグループは重複しており，重複のあるグループ単位で特徴選択するために，重複のあるグループノルムに基づくスパース回帰法が提案されている [Jacob2009]．グループ g に含まれる特徴量 $d \in \mathcal{G}_g$ に対応するパラメータ $\boldsymbol{v}_g$ とする．$\boldsymbol{v}_g$ は D 次元ベクトルであるが，グループに含まれない特徴量に対応する要素値をゼロと固定する．ロジスティック回帰モデルのパラメータ $\boldsymbol{w} = \sum_{g=1}^{G} \boldsymbol{v}_g$ としたとき，重複ありグループのノルムは

$$\Omega_{\text{overlap}}(\boldsymbol{w}; \boldsymbol{\alpha}) = \inf_{\sum_g \boldsymbol{v}_g = \boldsymbol{w}} \sum_{g=1}^{G} \alpha_g \cdot \|\boldsymbol{v}_g\|_2 \tag{4}$$

と定義され，全体のコスト関数は

$$J_{\text{group}}(\boldsymbol{w}) = L(\boldsymbol{w}) + \lambda \cdot \Omega_{\text{overlap}}(\boldsymbol{w}) \tag{5}$$

となる．ただし，λ は全体の正則化重みであり，α_g はグループ毎の正則化重みとなる．

本研究では，標本の部分サンプリングおよび特徴量の正則化重みのランダム化による特徴量の重要度の計算法を提案する．標本の部分サンプリングは外れ値と考えられる標本の影響を抑えるために行っており，また，特徴量の正則化重みのランダム化は影響の強い特徴量のみを取り出すために行っている．具体的な手順は，以下の通りである．

1. N 個の標本から r % の標本を無作為に選択する．
2. 範囲 $[\theta, 1]$ の一様乱数 η_g を用いて $\alpha_g = \eta_g$ とグループごとの正則化重みを決定する．

連絡先: 志賀元紀，岐阜大学 工学部 電気電子・情報工学科，shiga_m@gifu-u.ac.jp

3. 手順 1 と 2 で得られた値を用いて $\boldsymbol{w}$ を最適化し，$w_d > 0$ に対応する特徴量 d を選択する．

4. 手順 1–3 を反復し，選択された割合 p_d を特徴量 d の重要度とする．

ただし，λ, r, θ はユーザーが設定すべきハイパーパラメータである．乱数生成後のモデルパラメータ $\boldsymbol{w}$ の最適化には加速つき近接勾配法 [Back 2009] を用いた．

3. 数値実験

3.1 人工データ

目的変数と関係ある特徴量（有効特徴量）を 6 個，無関係な特徴量（雑音特徴量）を 30 個とした合計 36 個の特徴量からなる人工データを生成した．標本 n の目的変数 y_n をしたとき，有効特徴量を平均 y_n および分散 σ^2 の正規分布から生成し，雑音特徴量は目的変数と無関係に平均 0 および分散 1 の正規分布から生成した．生成されたデータを特徴量毎に標準化した後，特徴選択法および予測法に入力された．特徴量のグループはそれぞれ 4 つの特徴量を含むように仮定され，つまり，i 番目のグループを $\mathcal{G}_i = \{d_{j+2(i-1)} | j = 1, 2, 3, 4\}$, $i = 1, 2, \cdots, 17$ とした．正例および負例を等しい数として，30 個の訓練データおよび 1000 個のテストデータを生成して，以下の通り提案法の性能検証のための実験を行った．

正則化重みの様々な値に対して選択される特徴量を調べるために，$\sigma^2 = 0.75$ および $\sigma^2 = 1$ とした人工データの正則化パスを計算し，それぞれ図 1 と図 2 に示した．提案法（RGLR）の他に，比較のために L1 正則ロジスティック回帰（LR），Group LR (GLR)，Randomized LR (RLR) を実行した．RLR と提案法において，ハイパーパラメータを $(r, \theta) = (90, 0.1)$ と設定した．図の横軸は正則化重み λ の対数値であり，縦軸は特徴量の重要度（回帰係数の絶対値または選択された確率）である．また，有効特徴量を色付きの実線で示し，雑音特徴量を黒色の破線で示した．

図 1 と 2 において，方法 LR と GLR の比較および方法 RLR と RGLR の比較より，グループ情報を用いることによって正則化重みが大きい場合に，有効特徴量の重要度が雑音特徴量の重要度よりも概して大きい重要度を示している．また，方法 LR と RLR の比較および方法 GLR と RGLR の比較より，サブサンプリングおよび正則化重みのランダム化によって雑音特徴量より有効特徴量の重要度が大きくなることが分かる．これらの結果より，グループ補助情報およぶランダムネスの導入によって，提案法は優れた特徴選択を実現できることが示されている．また，RLR では有効特徴量と雑音特徴量との差が小さいため，閾値設定が難しい．提案法 RGLR の正則化パスにおいて有効特徴量の重要度がゼロになる重みの値が雑音特徴量における値と明確に区別しやすく，適切な重みを設定しやすいことが分かる．

次に，正則化重み $\log_{10}(\lambda) = -1.053$ と固定して特徴量の重要度ランキングを詳細に調べた．この値は，全ての特徴選択法において，全ての重要度が 0 にならず一部の特徴量が選択される値に決定した．有効特徴量を正例，雑音特徴量を負例として，重要度によるランキング結果から AUC (Area Under the ROC curve) を計算し，特徴選択の正確性の評価基準とした．異なる人工データを用いた実験を 50 回行い，AUC の平均値と標準偏差を表 1 に図示する．これらの結果より，提案法の AUC は他の結果より大きい値になることが分かり，提案法の特徴選択が優れることが示された．

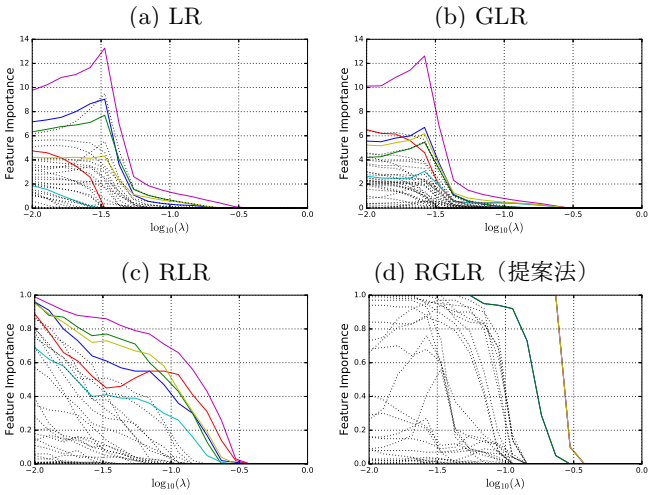

図 1: 人工データ ($\sigma^2 = 0.75$) の正則化パス

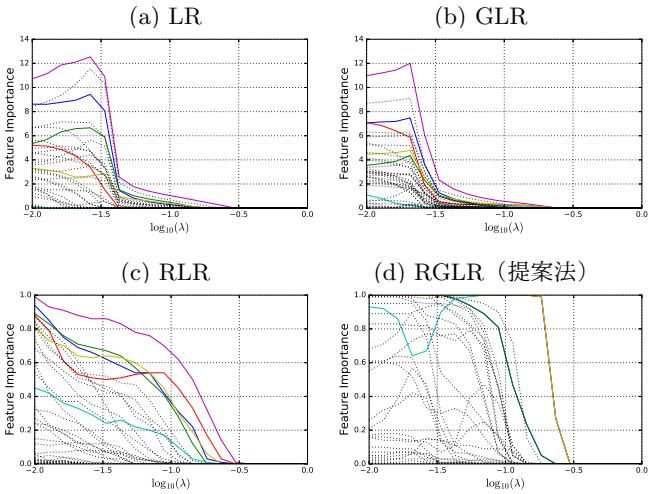

図 2: 人工データ ($\sigma^2 = 1.0$) の正則化パス

次に，テストデータにおける予測性能を比較した．前述の特徴選択法（正則化重み $\log_{10}(\lambda) = -1.053$）で選ばれた特徴を用いて，L2 正則化ロジスティック回帰を用いて予測した．特徴選択に用いた重要度の閾値を，LR と GLR の閾値 0 とし，RLR と RGLR（提案法）では閾値 θ_p を 0.25 または 0.5 の 2 通りを設定した．数値実験の 50 回繰り返して計算された AUC の平均値と標準偏差を表 2 に示した．表 2 より，分散 σ^2 がどちらの値の場合においても，提案法 RGLR の AUC が他の手法と比較して若干大きな値を示している．これらの実験結果から，提案法では有効特徴量を安定的に選択できているため，目的変数の予測性能が向上したと考えられる．

3.2 実データ

先行研究 [Zaas2009, Woods2013] で使用された 2 種類のインフルエンザ H1N1 および H3N2 に関するデータをを使用した．それぞれのデータセットに含まれる被験者数は，H1N1 が 24 人であり，H3N2 が 17 人である．各被験者の目的変数は曝露後数日間の感染症状スコアより導かれた 2 値変数とした．また，早期発見を目的とするため，ウイルスの曝露直前（時刻の閾値 $\theta_t \leq 0$）および曝露 1 日後（$\theta_t \leq 24$）の範囲の 2 種類のデータセットの取り出し，それぞれ，被験者および遺伝子毎に平均したデータを特徴量に用いた．本課題は遺伝子発現

特徴選択法	$\sigma^2 = 0.75$	$\sigma^2 = 1.0$
LR	0.91 ± 0.07	0.86 ± 0.08
GLR	$\mathbf{0.99 \pm 0.03}$	$\mathbf{0.96 \pm 0.08}$
RLR	0.98 ± 0.02	0.95 ± 0.04
RGLR	$\mathbf{0.99 \pm 0.02}$	$\mathbf{0.98 \pm 0.04}$

表 1: 特徴量のランキング性能 AUC

特徴選択法	$\sigma^2 = 0.75$	$\sigma^2 = 1.0$
LR	0.94 ± 0.03	0.85 ± 0.06
GLR	0.95 ± 0.03	0.86 ± 0.06
RLR ($\theta_p = 0.25$)	0.95 ± 0.03	$\mathbf{0.87 \pm 0.06}$
RLR ($\theta_p = 0.50$)	0.95 ± 0.03	0.85 ± 0.06
RGLR ($\theta_p = 0.25$)	$\mathbf{0.96 \pm 0.03}$	$\mathbf{0.87 \pm 0.06}$
RGLR ($\theta_p = 0.50$)	$\mathbf{0.96 \pm 0.03}$	$\mathbf{0.87 \pm 0.06}$

表 2: 人工データにおける予測性能 AUC

量から将来の感染症状悪化を予測する問題として定式化される．また，遺伝子のグループにはヒト遺伝疾患データベース OMIM [OMIM2017] を用いた．グループ補助情報の有効性を検証する実験であるため，いずれのグループにも属さない遺伝子を取り除いた．また，発現量の分散が大きい遺伝子のみを解析対象にした．結果として，データセットに含まれる遺伝子数は 100，また，グループ数は 40 となった．本実験では，まず，2 層 4 分割交差検証を用いて全ての被験者の陽性スコアを計算し，その後，全ての被験者のスコアによって AUC を計算した．本実験で用いた特徴選択法および予測手法は人工データ実験と同じ手法である．

正則化パスの一例として，H1N1 を曝露直前のデータに対する結果を図 3 に示す．実データの有効特徴量は未知であるため，全て黒色の実線で示した．図 3 の正則化パスの傾向は，人工データの正則化パス（図 1 と 2）と類似する形状を示し，正則化重み λ が大きい値を選択すれば，λ の値が多少変化しても選択される特徴量が変化しないことがわかる．

次に，正則化重みを $\log_{10}(\lambda) = -1.966$ に固定して特徴選択を行い，選択された特徴量のみを用いて予測した際の予測性能 AUC を表 3 に示した．H1N1 の曝露直前のデータセットにおいては，提案法の予測性能が比較法の予測性能より優れていることが分かる．しかしながら，他のデータセットでは従来法の予測性能が優れており，提案法による特徴選択の有効性を確認できなかった．全体的に，曝露直前および曝露 1 日後までのデータを用いた予測性能を比較したところ，曝露直前の予測結果の方が優れていることが分かる．これは，データを平均化処理することで曝露 1 日後の時系列変化の情報が取り除かれてしまった可能性がある．この問題を解決するために，時系列変化の違いを捉えるための特徴量の表現が必要とされる．

4. まとめ

本研究では，グループ補助情報およびランダムネスの導入に基づく安定的な特徴選択法を提案し，その性能を数値実験により検証した．数値実験の際に，正則化重みをマニュアルで設定したものの，実応用では自動設定が求められる．標本数が少ない場合，交差検証によって適切な重みを選択できないため，この問題点の解消が今後の課題として挙げられる．

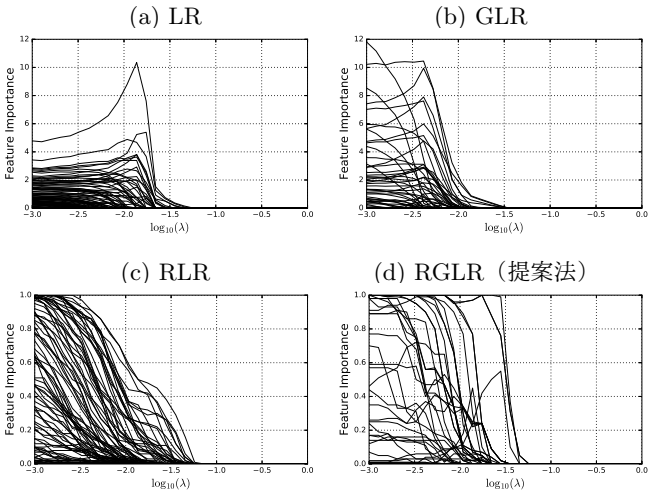

図 3: 実データ (Influenza H1N1, $\theta_t \leq 0$) の正則化パス

	H1N1		H3N2	
特徴選択法	$\theta_t \leq 0$	$\theta_t \leq 24$	$\theta_t \leq 0$	$\theta_t \leq 24$
LR	0.77	0.69	**0.82**	**0.76**
GLR	0.80	0.53	**0.92**	**0.72**
RLR ($\theta_p = 0.25$)	0.76	**0.82**	0.72	0.57
RLR ($\theta_p = 0.50$)	0.74	**0.84**	0.64	0.42
RGLR ($\theta_p = 0.25$)	**0.88**	0.69	0.81	0.49
RGLR ($\theta_p = 0.50$)	**0.91**	0.56	0.79	0.43

表 3: 実データにおける予測性能 AUC

参考文献

[Nicolai 2010] Nicolai, M. and Bühlmann, P., "Stability Selection," *Journal of the Royal Statistical Society B*, 72 (4), 417–473, (2010).

[Jacob2009] Jacob, L., Obozinski, G. and Vert, J.P., "Group lasso with overlap and graph lasso," *Proc. of the 26th ICML*, 433–440, 2009.

[Back 2009] Beck, A., and Teboulle, M., "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, 2(1), 183–202, 2009.

[Zaas2009] Zaas, A.K. and *et al.*, "Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans," *Cell Host and Microbe*, 6(3), 207–217, 2009.

[Woods2013] Woods, C.W. and *et al.*, "A Host Transcriptional Signature for Presymptomatic Detection of Infection in Humans Exposed to Influenza H1N1 or H3N2," *PLOS ONE*, 8(1), e52198, 2013.

[OMIM2017] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), March 7, 2017. (https://omim.org/)