G-021

ベイズ最適化と対照学習を導入したマルチオミクスデータ統合による疾患分類
Disease Classification Introducing Multi-Omics Data Integration with Bayesian Optimization and Contrastive Learning

鈴木 翔大 [1)]　佐久間 拓人[1)]　加藤 昇平 [1)]
Shodai Suzuki　Takuto Sakuma　Shohei Kato

1　はじめに

オミクス解析とは，遺伝子，mRNA，タンパク質，代謝物といった多様な分子に関する，生物系の複雑な情報を網羅的に取得・解析する手法である．また，一般的な疾患の病因は多因子的であり，これらの複数のオミクス解析を統合するマルチオミクスアプローチを用いることで，科学的根拠に基づいたより確実な結論を導き出すことができる．この手法は，特に代謝性疾患，神経疾患，がんなどの多因子疾患において，診断ツールの開発や新規治療標的の同定に有効である [1]．一方で，これらのデータは非常に複雑であり，機械学習による解析手法の研究が進められているものの，マルチオミクス解析においては，各オミクスデータをいかに効果的に統合するかが大きな課題となっている．こうした中で Wang ら [2] は，深層学習の一種であるグラフ畳み込みニューラルネットワーク（GCN:Graph Convolutional Neural Networks）と VCDN（View Correlation Discovery Network）という独自の統合方法を用いて患者分類を実施し，これによって各データ間の相関関係を利用することによって分類性能が向上する可能性を示唆した．一方で，この手法には，各オミクスの影響力を考慮できていない問題があると考えられる．また Yang ら [3] は，このタスクにおける対照学習の有効性を示唆している．対照学習とは，データ同士を比較する仕組みを用いて似たデータは近づき，異なるデータは遠ざかるように特徴量を学習する手法である．本研究では，Wang らのモデルにおける課題を解決するため，ベイズ最適化と対照学習を用いたモデルの開発・評価を実施する．

2　提案手法

2.1　データセット

本研究では，異なる 2 つのデータセットを使用して提案モデルの有効性を検証した．アルツハイマー病（AD）患者と正常対照（NC）の分類には AD Knowledge Portal[4] から取得した ROSMAP，乳がんのサブタイプ分類には TCGA から取得した BRCA[5] を使用した．これらのデータセットには mRNA 発現データ（mRNA），DNA メチル化データ（Meth），miRNA 発現データ（miRNA）の 3 種類のオミクスデータをもつサンプルが含まれている．各オミクスデータに対して Wang らと同様の前処理と特徴の事前選択を施した．表 1 と表 2 にそれぞれ ROSMAP と BRCA に含まれる各クラスのサンプル数と各オミクスの特徴数を示す．

2.2　提案モデル

Wang ら [4] の提案するマルチオミクス統合手法では，各オミクスの 5 クラス分類の結果から行列積を算出することによってオミクスデータを結合する．これによってオミクス間の相関関係を反映し，高い性能を発揮することが報告されている．一方で，各オミクスが分類結果に与える影響力は多様であり，Wang らのモデルではこうした影響力を反映できない．各オミクスの分類における重要度をモデルに反映させることで，さらなる分類性能の向上が期待できると考え，本研究ではベイズ最適化を導入してエンコーダの出力次元数を調整し，行列積による統合を通じて分類を実施するモデルを提案する．しかし，先行研究では各オミクスにおける分類結果が最終的な分類に有効であるとされていたが，ベイズ最適化によって出力次元数を調整することで，その情報が失われる恐れがある．それを防ぐため，本研究ではベイズ最適化に加え，エンコーダに対して対照学習を適用した．本稿では教師あり対照学習に注目し，ラベルごとに出力が遠ざかるように学習した．図 1 に提案するモデルの概要を示す．ここで各オミクスの出力次元数は学習ごとに変化し，例として，表 3 に BRCA に対して後述するモデル 4 を適用した際の出力次元数を示す．先行研究では 5 クラス分類の結果のため出力次元数はすべて 5 であったが，モデル 4 では mRNA>miRNA>Meth となっており，ベイズ最適化によってこれらが変化していることが確認できる．

表 1: ROSMAP の概要

Omics Type	Number of Features	Number After Preselection
mRNA	55889	200
DNA Methylation	23788	200
miRNA	309	200
Categories:		
normal control(NC): 169,　Alzheimer’s disease (AD): 182		

表 2: BRCA の概要

Omics Type	Number of Features	Number After Preselection
mRNA	20531	1000
DNA Methylation	20106	1000
miRNA	503	503
Categories:		
Normal-like: 115,　Basal-like: 131,　HER2-enriched: 46,　Luminal A: 436,　Luminal B: 147		

3　実験と評価

3.1　比較モデル

今回はベイズ最適化と対照学習，またそれを組合せたときの有用性をそれぞれ調査するために，以下の 4 つのモデルを用意した．

1.先行研究と同様に統合したモデル 1

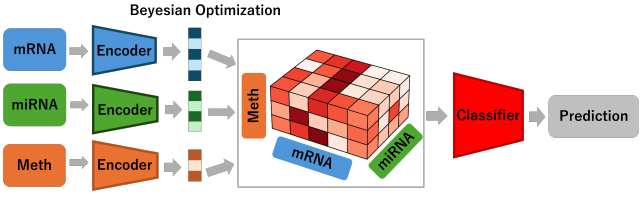

図 1: モデルの概要

1）　名古屋工業大学 大学院工学研究科 工学専攻
Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology

表 3: 各オミクスの出力次元数（BRCA）

Omics	Output Dimensions
mRNA	33.52 ± 10.46
miRNA	18.40 ± 13.56
Meth	14.50 ± 12.31

表 4: 各モデルの Accuracy および F1 スコア（ROSMAP）

Model	Accuracy	Macro-Average F1 Score
Model 1	0.655 ± 0.108	0.637 ± 0.128
Model 2	0.679 ± 0.058	0.624 ± 0.130
Model 3	0.674 ± 0.101	0.657 ± 0.126
Model 4	**0.744 ± 0.051**	**0.744 ± 0.075**

2.ベイズ最適化により出力次元数を調整したモデル 2

3.エンコーダに対照学習を採用したモデル 3

4.ベイズ最適化と対照学習両方を採用したモデル 4

この時，モデル 1 およびモデル 2 のエンコーダについては 3 層の全結合層からなる全結合ニューラルネットを採用した．実験は 50 試行実施し，各試行では，データを訓練用・検証用・テスト用として 6:2:2 の割合でランダムに分割した．モデルを訓練データによる学習と，検証データによる早期停止に基づいて構築した後，テストデータを用いて最終的な分類を実施した．評価指標としては，ROSMAP では Accuracy と F1 スコア，BRCA では Accuracy とマクロ平均 F1 スコアを用いた．いずれも有意水準 0.05 で両側 t 検定を実施し，Holm 法による多重検定補正を適用した．

3.2 結果

ROSMAP における実験結果について，表 4 に各モデルの Accuracy および F1 スコアを，表 5 に両側 t 検定による p 値をそれぞれ示す．どちらの指標においてもモデル 4 が最も高い性能を示した．また，t 検定の結果，モデル 4 とそれ以外のモデルの間に有意な差が確認された．同様に，BRCA における実験結果は表 6 に Accuracy およびマクロ平均 F1 スコアを，表 7 に両側 t 検定による p 値をそれぞれ示す．BRCA においても，両指標においてモデル 4 が最も高い性能を示した．また，t 検定の結果，モデル 2〜4 はいずれもモデル 1 に対して有意な差が確認された．

3.3 考察

実験結果から，提案手法であるモデル 4 がモデル 1 に対して有意に改善し，先行研究における統合方法よりも有用であることが示唆された．BRCA の実験ではモデル 2 とモデル 3 がモデル 1 に対して有意に改善したことから，ベイズ最適化と対照学習それぞれの有用性が示唆された．また，ROSMAP の実験では，モデル 2 およびモデル 3 がモデル 1 に対して有意な差を示さなかったのに対し，モデル 4 はモデル 1 に対して有意な差を示しており，ベイズ最適化と対照学習を組み合わせることによる有用性が示唆された．一方で，BRCA の実験においてモデル 4 とモデル 2 およびモデル 3 との間にあまり有意差が見られなかったことから，5 クラス分類を対象とする BRCA では対照学習による効果が大きく，対照学習のみでも十分な効果を発揮した可能性がある．

表 5: 両側 t 検定による p 値（ROSMAP)

Model A	Model B	p-value (Accuracy)	p-value (F1)
Model 1	Model 2	0.460	0.617
Model 1	Model 3	0.605	0.614
Model 1	Model 4	$\mathbf{1.8 \times 10^{-5}}$	$\mathbf{3.3 \times 10^{-5}}$
Model 2	Model 3	0.786	0.570
Model 2	Model 4	$\mathbf{2.9 \times 10^{-7}}$	$\mathbf{1.7 \times 10^{-6}}$
Model 3	Model 4	$\mathbf{1.7 \times 10^{-4}}$	$\mathbf{1.1 \times 10^{-4}}$

表 6: 各モデルの Accuracy および F1 スコア（BRCA）

Model	Accuracy	Macro-Average F1 Score
Model 1	0.795 ± 0.027	0.773 ± 0.036
Model 2	0.805 ± 0.030	0.792 ± 0.034
Model 3	0.819 ± 0.025	0.790 ± 0.039
Model 4	**0.824 ± 0.035**	**0.798 ± 0.048**

表 7: 両側 t 検定による p 値（BRCA)

Model A	Model B	p-value (Accuracy)	p-value (Macro F1)
Model 1	Model 2	$\mathbf{2.1 \times 10^{-2}}$	$\mathbf{9.0 \times 10^{-5}}$
Model 1	Model 3	$\mathbf{1.9 \times 10^{-7}}$	$\mathbf{1.7 \times 10^{-4}}$
Model 1	Model 4	$\mathbf{3.3 \times 10^{-7}}$	$\mathbf{1.7 \times 10^{-4}}$
Model 2	Model 3	$\mathbf{7.2 \times 10^{-3}}$	7.9×10^{-1}
Model 2	Model 4	$\mathbf{2.8 \times 10^{-3}}$	7.9×10^{-1}
Model 3	Model 4	2.0×10^{-1}	6.6×10^{-1}

4 まとめと今後の課題

本研究では，ベイズ最適化と対照学習を組合せた新たなマルチオミクス統合モデルを提案した．実験の結果，提案モデルは 2 つのデータセットにおいて先行研究における統合モデルと比較して，有意差 0.05 の両側 t 検定において有意な差を獲得した．今回は本モデルの有用性をシンプルな構造で検証するために，エンコーダに全結合層のみを採用したが，今後は様々なエンコーダを用いてさらに精度を高めていきたい．また，将来的には直接的な予後の予測やバイオマーカーの同定を実施するモデルの構築に取り組みたい．

謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 JP24H00741），ならびに，国立研究開発法人情報通信研究機構委託研究の助成により行われた．

参考文献

[1] N. Perakakis, A. Yazdani, G. E. Karniadakis, and C. Mantzoros, ”Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics,” *Metabolism*, vol. 87, pp. A1–A9, 2018.

[2] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, ”MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification,” *Nat. Commun.*, vol. 12, p. 3445, 2021.

[3] M. Yang, Y. Yang, C. Xie, M. Ni, J. Liu, H. Yang, F. Mu, and J. Wang, ”Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale,” *Nat. Mach. Intell.*, vol. 4, pp. 696–709, 2022.

[4] Hodes, R. J. & Buckholtz, N. Accelerating medicines partnership: Alzheimer’s disease (amp-ad) knowledge portal aids alzheimer’s drug discovery through open data sharing. Expert Opin. Ther. Tar. 20, 389–391 (2016)

[5] National Cancer Institute at the National Institutes of Health, ”TCGA-BRCA,” [Online]. Available: `https://portal.gdc.cancer.gov/projects/TCGA-BRCA` (accessed: Jan. 25, 2025).