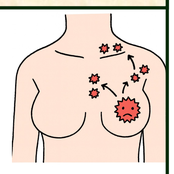
乳癌におけるGraph-CNN を用いた転移予測モデル

明治大学 総合数理学 現象数理学科 池田研究室

1. 背景

乳癌は女性で最も頻度の高い癌の一つであり、

乳癌の治療成績を左右する最大の因子は転移の有無



従来の機械学習モデルは、遺伝子を独立した特徴量としてしか認識できない

従来の方法

遺伝子A	1.23
遺伝子B	0.87
遺伝子C	2.54

→ ML Model

遺伝子A　遺伝子A

“遺伝子同士の”がモデルに反映されない

従来の機械学習モデルでは

“遺伝子同士の相互作用（ネットワーク構造）”を十分に扱えない

遺伝子同士の関連性がわかる PPIネットワーク → 組み込む

CNNをグラフ構造を持つデータに対して行うように拡張したモデル

Graph-CNNが有効

2. 目標

Graph-CNNを用いて乳癌の転移を予測する

3. 先行研究と手法

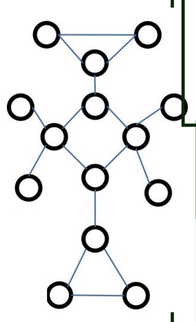
一次データ

遺伝子発現データ https://toil.xenahubs.net/download/tcga_RSEM_gene_tpm.gz

臨床データ https://www.cbioportal.org/study/clinicalData?id=brca_tcga_pan_can_atlas_2018

患者

PPIネットワーク（二次データ）



遺伝子データ（一次データ）

隣接行列（二次データ）

$$A_{ij} = \begin{cases} 1 & \text{遺伝子iとjに相互作用有} \\ 0 & \text{それ以外} \end{cases}$$

次数行列

$$D_{ii} = \sum_j A_{ij}$$

グラフラプラシアン

$L = D - A$ の固有値$\lambda \in [0, \lambda_{max}]$

↓ 固有値を-1から1に収めるため

$\tilde{L} = \frac{2L}{\lambda_{max}} - I$ に変換して、

固有値μをに写像させる

Chebyshev多項式は入力値の範囲を[-1, 1]にすることで安定するため、代入する前に調整を行なった。

Chebyshev多項式

$$T_0(\tilde{L}) = I \qquad T_1(\tilde{L}) = \tilde{L}$$

$$T_k(\tilde{L}) = 2\tilde{L}T_{k-1}(\tilde{L}) - T_{k-2}(\tilde{L})$$

次数Kの意味：K hop先までのノードをフィルタする

Q. なぜこの多項式を使用したか？

A. 自然にK-hop近傍を表現できるから

Kを自分で定めることで、モデルのノードの影響範囲を自分で調整できる

※本研究ではK=7

ただし、実際のChebyshev畳み込みは計算の高速化を測るために

固有値分解はせずに、$T_k(\tilde{L})$の再帰計算だけ使用する

畳み込み本体

$$y = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) x$$

θ_k:学習されるパラメータ

$T_k(\tilde{L})$:近傍情報を含んだフィルタ

x:ある患者の遺伝子発現ベクトル

↓

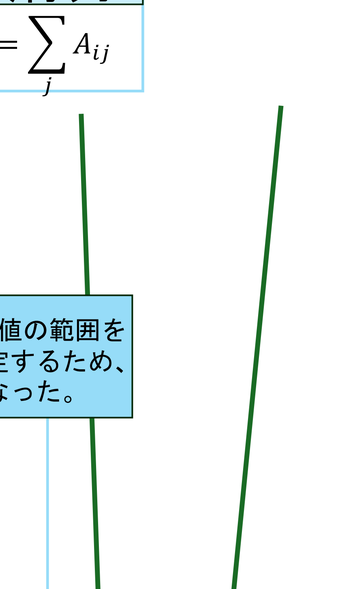
-多層構造-

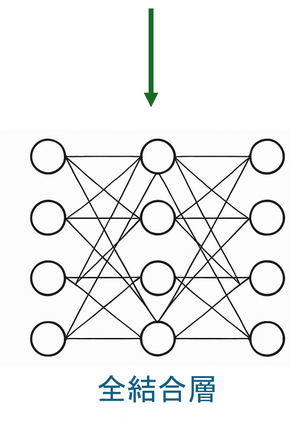
F個のフィルタがある時

$$y^{(j)} = \sum_{k=0}^{K-1} \theta_k{}^{(j)} T_k(\tilde{L}^{(j)}) x^{(j)} (\mathrm{j} = 1, \ldots, \mathrm{F})$$

$y^{(j)}$：j番目のフィルタの出力

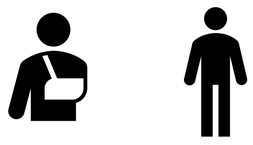
畳み込み





全結合層

転移あり　転移なし



MLPの役割は
重要な特徴を強調
不要な特徴を抑制

全結合層

yの各要素をどれだけ重要視するかを学習する

アフィン変換

$$h_1 = ReLU(W_1 y + b_1)$$

ReLU(x)=max(0, x) →ReLUがあること で非線形な関係も学習が可能になる

h_1の情報をまた重み付けして組み合わせる
分類に必要な最終的な特徴を形成

$$h_2 = ReLU(W_2 h_1 + b_2)$$

(W_1, W_2:重み(行列)　y:入力ベクトル
b_1, b_2:バイアス　(畳み込みで出力されたベクトル))

出力

$$\hat{y} = \sigma(W h_2 + b)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$\hat{y} \approx 0$　転移なし

$\hat{y} \approx 1$　転移あり

W:最終層の重み(学習される)

b:バイアス(学習される)

h_2:MLPの最後の隠れ層の特徴

σ:シグモイド関数

クラス不均衡の問題について

臨床データ(転移ありor転移なし)のデータ数が不均衡

↓

class_weightを使用し、重みを調整

少数クラスを誤分類した時のペナルティを大きくする

→本研究では　0:0.7125倍　1:1.676倍に設定

4. 評価指標

損失関数 Binary Cross Entropy

$$\mathcal{L}_{train} = -\frac{1}{N_{train}} \sum_{i \in train}^{N} [y_i log\hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]$$

$y_i \epsilon \{0,1\}$ 正解ラベル
$\hat{y}_i = \sigma(z_i)$ モデルの出力確率

$$\mathcal{L}_{val} = -\frac{1}{N_{val}} \sum_{i \in val}^{N} [y_i log\hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]$$

$\mathcal{L}_{train}$の結果に基づいて、値が最小になるように重みを調整し、
モデルの精度向上を試みる。

$\mathcal{L}_{val}$の値は過学習したタイミングで上がり始める。

5. 結果

出力結果

	Train loss	Val loss
1回目	0.7424	0.6082
21回目	0.6833	0.7305

減少　増加

Train loss:訓練データの誤差　重みを更新するために使用

Val loss:検証データで測る誤差　過学習を確認するために使用

学習回数:21回(過学習が起こったため、途中で停止)
Train lossは減少したものの、Val lossが過学習により増加してしまった。
→良い精度とは言えない。

6. 結論

精度は低いものの、乳癌の転移を予測できた。

	Train loss	Val loss