修士論文 要旨

Graphon を用いた Graph Convolutional Neural Network の推定理論

数理情報学専攻 48206206　大澤 龍太
指導教員　鈴木 大慈 准教授

1　はじめに

近年，グラフ上の深層学習である Graph Neural Networks (GNN) の標準的な一種である Graph Convolutional Networks (GCN) [1] は，広範なグラフ機械学習タスクにおいて，最先端の結果を達成した．GCN の特徴的な点は畳み込みと呼ばれる操作にある．畳み込みとは，グラフ上の各ノードの特徴量を，その近傍にあるノードの特徴量と平均化する操作である．これは直感的には，特徴量をグラフの隣接関係によって決まる滑らかさのもとで平滑化する操作であると解釈できるが，エッジがノイズを含んでいたり，未観測のノードが作用したりするとき，GCN がグラフの背後にあると仮定される「真の滑らかさ」を復元できるかは明らかではない．本研究では Graphon [2] によるランダムグラフモデルのもとで，グラフの「真の滑らかさ」を仮定し，真の滑らかさを反映した信号の復元の統計的推定誤差という観点から，GCN とその変種である GCNII [3] の学習能力を検討する．さらに，得られた結果を類似の問題であるカーネル法による関数推定問題と比較する．

2　既存研究

2.1　GCN とその変種

GCN の ℓ 層目の更新は，次式で定義される:

$$X^{(\ell+1)} = \sigma\left(\hat{A}X^{(\ell)}W^{(\ell)}\right).$$

ここで，$\hat{A} \in \mathbb{R}^{n\times n}$ は正規化隣接行列であり，$X^{(\ell)} \in \mathbb{R}^{n\times d_\ell}$ は ℓ 層目の特徴行列である．また，$W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell+1}}$ は重み行列である．GCN の特徴的な点は畳み込みと呼ばれる，隣接行列によって特徴量を平滑化する操作にあるが，GCN は畳み込みを繰り返すことで，ノードの表現が均一化される over-smoothing と呼ばれる状態に陥る．この問題を受けて，[3] が提案した GCNII では GCN の更新式において，以下の 2 つの置き換えを行い，over-smoothing に対処する:

$$\begin{aligned}\hat{A}X^{(l)} &\leftarrow (1-\alpha^{(\ell)})\hat{A}X^{(\ell)} + \alpha^{(\ell)}X^{(0)},\\ W^{(\ell)} &\leftarrow \left(1-\beta^{(\ell)}\right)I + \beta^{(\ell)}W^{(\ell)}.\end{aligned}$$

2.2　Graphon

確率空間 $(\mathcal{U}, \mathcal{F}_\mathcal{U}, P_\mathcal{U})$ を考える．Graphon [2] は対称な可測関数 $W : \mathcal{U}^2 \to [0,1]$ であり，積分作用素 $\mathbb{W} : L^2(\mathcal{U}) \ni f \mapsto \int_\mathcal{U} W(\cdot, v)f(v)\mathrm{d}P_\mathcal{U}(v) \in L^2(\mathcal{U})$ を定義する．$\mathbb{W}$ は可算無限個の実固有値 $\lambda(\mathbb{W}) := \{\lambda_i\}_{i=1}^\infty$ を持つ．ここで，$\lambda(\mathbb{W})$ は絶対値について降順とする．

Graphon は，ランダムグラフの生成モデルとして用いられる．Graphon W からノード数 n のグラフ G をサンプリングする手順について述べる．まず，ノード v_i に関する潜在変数 $\mathcal{U} \ni u_i \sim P_\mathcal{U}$ のサンプリングを行う．次に，$\{u_i\}_{i=1}^n$ からエッジの存在確率を表す確率行列 $W \in \mathbb{R}^{n\times n}$ を構成する:

$$W_{ij} = W\left(u_i, u_j\right).$$

最後に，隣接行列 $A \in \mathbb{R}^{n\times n}$ のサンプリングを行う:

$$\begin{aligned}&A_{ij} \sim \mathrm{Ber}\left(\alpha_n W_{ij}\right) \quad (1 \le i < j \le n),\\ &A_{ii} = 0,\\ &A_{ij} = A_{ji}.\end{aligned}$$

ここで，Ber はベルヌーイ分布であり，$\alpha_n \in (0,1]$ はグラフのスパース性を制御するパラメータである．

3　GCN の推定

2.2 節で述べた手順に従って，潜在変数 $\{u_i\}_{i=1}^n \subset \mathcal{U}^n$ とグラフ G をサンプリングする．各ノード v_i について，$f : \mathcal{U} \to \mathbb{R}$ は有界関数として，特徴量 $f(u_i)$ と観測値

$$y_i = \mathbb{W}f(u_i) + \epsilon_i$$

が定まる．ここで，ϵ_i は独立で平均 0，分散 σ^2 のノイズである．目標は，グラフ G の情報を表す正規化隣接行列 $\bar{A} = \frac{1}{n\alpha_n}A$ と観測値 $y = [y_1, \ldots, y_n]^\top$ を用いた元信号 $\{\mathbb{W}f\}_{i=1}^n$ の復元である．復元は，線形活性化関数を持つ 1 層 GCN: $\hat{Y}_{\mathrm{GCN}} = \bar{A}h$ によって行い，$h \in \mathbb{R}^n$ を学習する．学習は，正則化項付きの平均二乗誤差の最小化によって行う:

$$\hat{h} = \arg\min_h \left\|y - \bar{A}h\right\|_2^2 + \lambda\left\|h\right\|_2^2 = (\bar{A}^2 + \lambda I)^{-1}\bar{A}y.$$

再構成された信号 $\bar{A}\hat{h}$ の性能を予測誤差 $\mathbb{E}_\epsilon\left[\frac{1}{n}\left\|g - \bar{A}\hat{h}\right\|_2^2\right]$ によって評価する．ここで，$g =$

$[\mathbb{W}f(u_1),\dots,\mathbb{W}f(u_n)]^\top$ とした．また，$\mathbb{E}_\epsilon$ は，潜在変数 $\{u_i\}_{i=1}^n$ に関して条件付けた条件付き期待値であり，ノイズ $\epsilon=[\epsilon_1,\dots,\epsilon_n]^\top$ に関して取る．

定理 1. 正則化項付き最小二乗推定量の平均二乗誤差は，以下のように抑えられる:

$$\mathbb{E}_\epsilon\left[\frac{1}{n}\|g-\bar{A}\hat{h}\|_2^2\right]\lesssim\lambda+\frac{N(\lambda)}{n}+\frac{1}{(n\alpha_n\lambda)^2}$$

が高確率で成り立つ．ここで，確率は潜在変数に関する確率である．また $N(\lambda)$ は $\mathbb{W}^2$ の自由度で，$N(\lambda):=\mathrm{Tr}\big((\mathbb{W}^2+\lambda\mathbb{I})^{-1}\mathbb{W}^2\big)$ である．

右辺第一項はバイアスの評価に由来し，第二項，第三項はバリアンスの評価に由来する．$\mathbb{W}$ の固有値 $\{\lambda_i\}_{i=1}^\infty$ について，多項式減衰 $|\lambda_i|\lesssim i^{-\gamma}$ を仮定すると，$N(\lambda)\lesssim\lambda^{-\frac{1}{2}}$ であり，バリアンスの支配項は $(n\alpha_n\lambda)^{-2}$ である．定理 1 をカーネル法の推定理論 [4] と比較すると，バイアスについてはカーネル法と類似する一方，バリアンスについては，$\frac{N(\lambda)}{n}$ が支配項とならず，隣接行列と確率行列の推定誤差に由来する第三項が支配項となるという点で異なる．

4 GCNII の推定

確率空間 $(\mathcal{U},\mathcal{F}_\mathcal{U},P_\mathcal{U}),(\mathcal{X},\mathcal{F}_\mathcal{X},P_\mathcal{X})$ を考える．第 3 章の手順とは異なり，各ノード v_i について，$f_1:\mathcal{U}\to\mathbb{R}$ は有界関数，$x_i\in\mathcal{X}$ は分布 $P_\mathcal{X}$ に従う確率変数として，特徴量 $\{f_1(u_i),\,x_i\}\subset\mathbb{R}\times\mathcal{X}$ と観測値

$$y_i=\mathbb{W}f_1(u_i)+\mathbb{K}^{\frac{1}{2}}f_2(x_i)+\epsilon_i$$

が定まる．ここで，$\mathbb{K}:L^2(\mathcal{X})\to L^2(\mathcal{X})$ は正定値カーネル $k:\mathcal{X}^2\to\mathbb{R}$ によって定まる積分作用素 $(\mathbb{K}f)(x):=\int_\mathcal{X}k(x,y)f(y)\mathrm{d}P_\mathcal{X}(y)$ であり，$f_2:\mathcal{X}\to\mathbb{R}$ は有界関数である．目標は，グラフ G の情報を表す正規化隣接行列 $\bar{A}=\frac{1}{n\alpha_n}A$ と特徴量 $\{x_i\}_{i=1}^n$，観測値 y を用いた元信号 $\{\mathbb{W}f_1(u_i)+\mathbb{K}^{\frac{1}{2}}f_2(x_i)\}_{i=1}^n$ の復元である．復元は，線形活性化関数を持つ 1 層 GCNII: $\hat{Y}_{\mathrm{GCNII}}=\bar{A}h+[F(x_1),\dots,F(x_n)]^\top$ によって行う．ここで，$h\in\mathbb{R}^n,F\in\mathcal{H}$ である．リプレゼンター定理から，$K:=[k(x_i,x_j)]$ として，学習は，以下の形の正則化項付きの平均二乗誤差の最小化によって行えばよい:

$$\hat{h},\hat{\alpha}=\frac{1}{n}\operatorname*{arg\,min}_{h,\alpha}\left\|\bar{A}h+K\alpha-y\right\|_2^2+\frac{\lambda_1}{n}\|h\|_2^2+\lambda_2\alpha^\top K\alpha,$$

$$\hat{h}=\lambda_2\bar{A}\Sigma^{-1}y,\quad\hat{\alpha}=\frac{\lambda_1}{n}\Sigma^{-1}y,$$

$$\Sigma:=\lambda_2\bar{A}^2+\lambda_1\bar{K}+\lambda_1\lambda_2 I\quad(\bar{K}:=\frac{1}{n}K).$$

再構成された信号 $\bar{A}\hat{h}+K\hat{\alpha}$ の性能を予測誤差 $\mathbb{E}_\epsilon\left[\frac{1}{n}\left\|g_1+g_2-\bar{A}\hat{h}-K\hat{\alpha}\right\|_2^2\right]$ によって評価する．ここで，$g_1=[\mathbb{W}f_1(u_1),\dots,\mathbb{W}f_1(u_n)]^\top,g_2=\left[\mathbb{K}^{\frac{1}{2}}f_2(x_1),\dots,\mathbb{K}^{\frac{1}{2}}f_2(x_n)\right]^\top$ とした．

定理 2.

$$\mathbb{E}_\epsilon\left[\frac{1}{n}\left\|g_1+g_2-\bar{A}\hat{h}-K\hat{\alpha}\right\|_2^2\right]\lesssim\lambda_1+\frac{N(\lambda_1)}{n}+\frac{1}{(n\alpha_n\lambda_1)^2}+\lambda_2+\frac{M(\lambda_2)}{n}$$

が高確率で成り立つ．

右辺第一項と第二項が $\bar{A}\hat{h}$ に関する結果であり，第三項と第四項が $K\hat{\alpha}$ に関する結果である．作用素 $\mathbb{K}$ の固有値 $\{\mu_i\}_{i=1}^\infty$ について，多項式減衰 $|\mu_i|\lesssim i^{-s}$ を仮定すると，$M(\lambda_2)\lesssim\lambda^{-\frac{1}{s}}$ である．適当な λ_1,λ_2 を選択することで，バイアスとバリアンスをバランスし，最適な収束レートを導出できる．

系 3. 定理 2 において，$\lambda_1\sim(n\alpha_n)^{-\frac{2}{3}},\lambda_2\sim n^{-\frac{s}{s+1}}$ とすることで，高確率で

$$\mathbb{E}_\epsilon\left[\frac{1}{n}\left\|g_1+g_2-\bar{A}\hat{h}-K\hat{\alpha}\right\|_2^2\right]\lesssim(n\alpha_n)^{-\frac{2}{3}}+n^{-\frac{s}{s+1}}$$

が成り立つ．

GCN とカーネル法の組み合わせとしての GCNII は，GCN によってグラフの隣接関係による滑らかさに基づく信号復元を行うと同時に，カーネル法によって特徴量に基づく関数推定を行う．全体の収束レートは，グラフのスパース性を表す α_n と，作用素 $\mathbb{K}$ の固有値の減衰速度 s の組み合わせによって決まり，ほとんどの場合において，GCN による信号復元が律速となるが，グラフが密 $(\alpha_n=1)$ で再生核ヒルベルト空間が複雑 (s が 1 に近い) な場合など，カーネル法による関数推定が律速となる場合もありうる．

参考文献

[1] Kipf, T. N., & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[2] Lovász, L. *Large Networks and Graph Limits* (Vol. 60). American Mathematical Society, 2012.

[3] Chen, M., Wei, Z., Huang, Z., Ding, B., & Li, Y. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, 2020, pp. 1725–1735.

[4] Caponnetto, A., & De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3), 331–368, 2007.