AcademiX

GNN勉強会

第5回 GNN勉強会

様々なタスクの応用4.1, 4.2章

山口大学大学院
創成科学研究科　平井元基

2024/08/04

https://www.academix.jp/

今回のゴール

今回のゴール
- GNNにおけるグラフ分類問題を定量的かつ定性的に理解をする。

学習手順
今回は、以下の図 1に示すような 3ステップでの学習を行います。

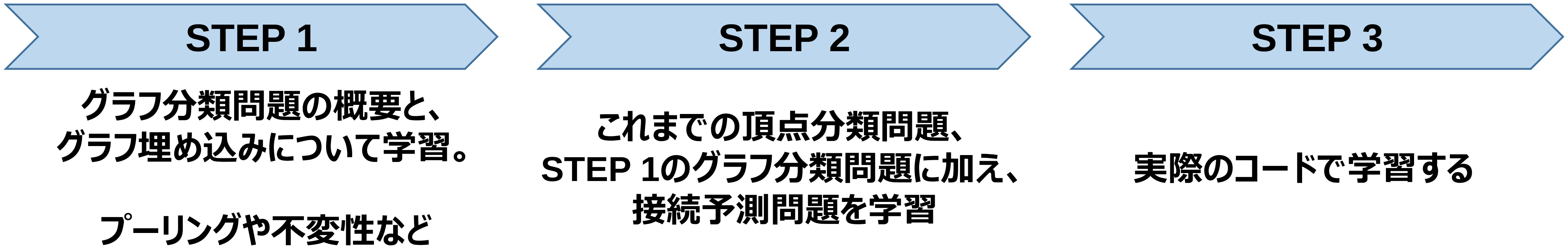


図1 第5回GNN勉強会におけるフロー図

参考書

今回使用するテキスト

佐藤 竜馬, “グラフニューラルネットワーク,”
講談社, 2024, 5月, 第1版, 第2印.
https://amzn.asia/d/hMovaf4

各種リンク

GitHub リポジトリ :
https://github.com/joisino/gnnbook.git

先生の自己紹介サイト
https://joisino.net/

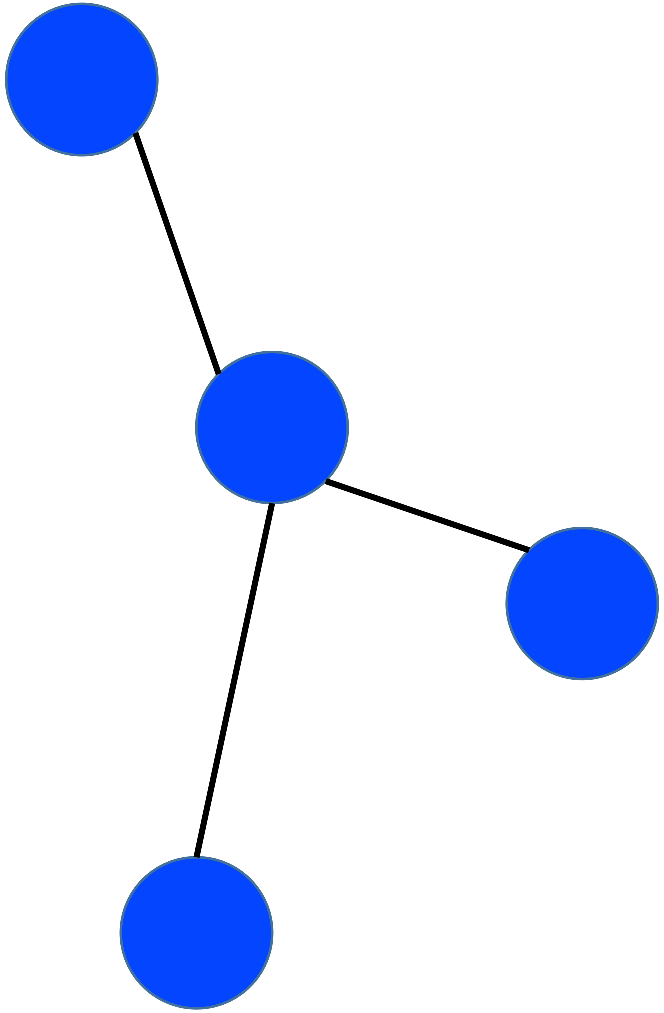
STEP 1

4.1 グラフ分類

4.1.1 グラフ分類問題の解き方

グラフ分類問題とは

- グラフ全体をクラスごとに分類する問題のこと。
  （例：化合物の毒性や効果の有無で分類、
  感情分析、画像分類、など）



具体的手法

- 大きく二つの手法が存在。
  ① グラフプーリング（Graph Pooling）を用いる方式
  ② 超頂点（supernode）を用いる方式

4.1.1 グラフ分類問題の解き方 -2

頂点分類問題とグラフ分類問題の比較

表1 頂点分類問題とグラフ分類問題の比較

GNNのタスク	頂点分類問題	グラフ分類問題
概要	グラフノードの分類	グラフそのものを分類
応用例	ユーザ分類、 画像のセグメンテーションなど	毒性予測、 画像分類など
学習手法	転導的学習、 帰納的学習	グラフプーリングの使用、 超頂点の使用

4.1.1 グラフ分類問題の解き方 -3

グラフプーリング

表2 グラフプーリング手法の比較

プーリングの種類	プーリングにおける数式
平均プーリング	$\boldsymbol{z}_G = \frac{1}{n}\sum_{v \in V} \boldsymbol{z}_v$
最大プーリング	$\boldsymbol{z}_G = max(\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n)$
頂点ごとに 重みを調整	$\boldsymbol{z}_G = \tanh\left(\sum_{v \in V} \sigma(f(\boldsymbol{z}_v, \boldsymbol{x}_v)) \odot \tanh(g(\boldsymbol{z}_v, \boldsymbol{x}_v))\right)$
Transformer	$\boldsymbol{z}_G = Transformer(\{\{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n, [CLS]\}\})[CLS]$

4.1.1 グラフ分類問題の解き方 -4

超頂点（supernode）とは

- 図 1のように、グラフの前処理手法の一つで、グラフに対し便宜的に導入するノードのこと。
- グラフ内のすべてのノードの情報を考慮して、一つのノードとして埋め込む。

超頂点の利点

- ノード数が異なるグラフ間の比較が容易となる。

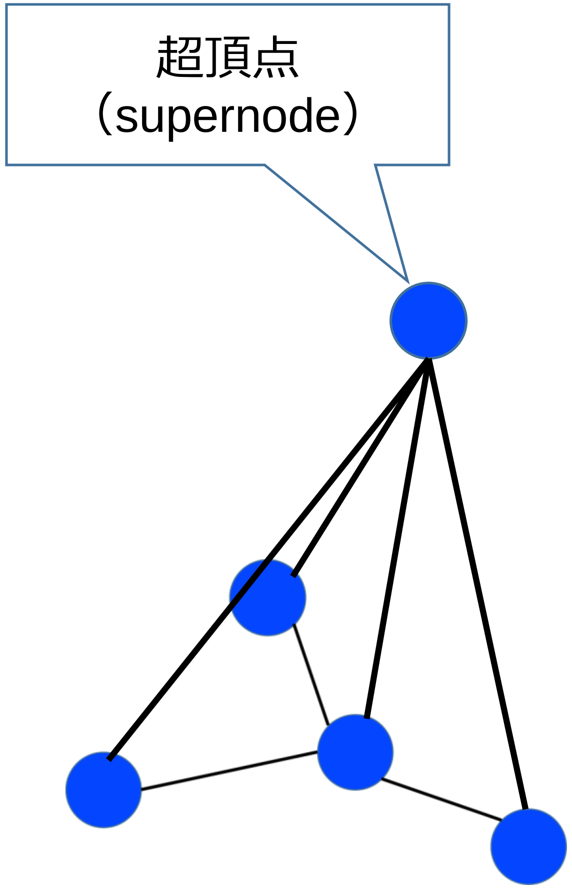


図2 超頂点の概略図

4.1.2 不変性

不変（invariant）とは

- グラフ埋め込み手法が不変（invariant）とは、
  同型グラフに対し、同じグラフ埋め込み手法を行うこと。
- 同変性（3.5章参照）の埋め込み版。

定義 4.2 （不変性）

グラフを受けとりグラフ埋め込み集合を返す関数

$$g : (V, E, \boldsymbol{X}) \mapsto \boldsymbol{z} \in \mathbb{R}^{\boldsymbol{d}}$$

が不変であることは、任意の同型なグラフ $G_1 = (V_1, E_1, \boldsymbol{X})$ と $G_2 = (V_2, E_2, \boldsymbol{Y})$ について、

$$g(G_1) = g(G_2)$$

が成り立つことをいう。
ただし、V を頂点集合、E を辺集合、$\boldsymbol{X}, \boldsymbol{Y}$ を頂点特徴量、$\boldsymbol{z}$ を埋め込みベクトルとする。

4.1.2 不変性 -2

不変を意識したプーリング（埋め込み）のメリット

- 解釈が容易
  （どのように予測したかが、直感的にわかりやすい）
- よい帰納バイアスをもつ
  （汎化性能が向上しやすい）

不変性のないプーリング手法

LSTMは系列の順番が重要なため、不変性をもたない

STEP 2

4.2 接続予測

4.2 接続予測

これまでの問題と接続予測問題の比較

表3 頂点分類問題とグラフ分類問題、および接続予測問題の比較

GNNのタスク	頂点分類問題	グラフ分類問題	接続予測問題
概要	グラフノードの分類	グラフそのものを分類	グラフの辺（エッジ）を予測
応用例	ユーザ分類、 画像のセグメンテーションなど	毒性予測、 画像分類など	推薦システム、 （新規の）ユーザ予測、など
学習手法	転導的学習、 帰納的学習	グラフプーリングの使用、 超頂点の使用	頂点埋め込みの対の使用、 グラフ分類に帰着する手法

4.2 接続予測 -2

頂点埋め込みの対を用いる方法

- まず頂点埋め込み $\boldsymbol{z}_v, \boldsymbol{z}_u$ を求め、埋め込みをもとに辺の有無を予測

具体的な学習手法

【例】

埋め込みの内積 $\boldsymbol{z}_v^\top \boldsymbol{z}_u$ が最も大きいときに辺が存在すると予測する手法

① 訓練時にはシグモイド関数を用い、$\sigma(\boldsymbol{z}_v^\top \boldsymbol{z}_u) \in (0, 1)$ を辺の存在確率とする。

② クロスエントロピー損失で学習。

③ テストデータにおける埋め込みベクトルの内積から、二値分類的に辺の有無を予測。

4.2 接続予測 -3

グラフ分類に帰着する方法

- 予測したい頂点の対 u, v について、u, v の周りのグラフを切り出し、切り出したグラフを正例・負例として分類して辺の有無を予測

具体的な学習手法

【例】

u, v のいずれかから K 本以下の辺を用いてグラフを切り出す手法
（図4.2[テキストより引用] 参考）

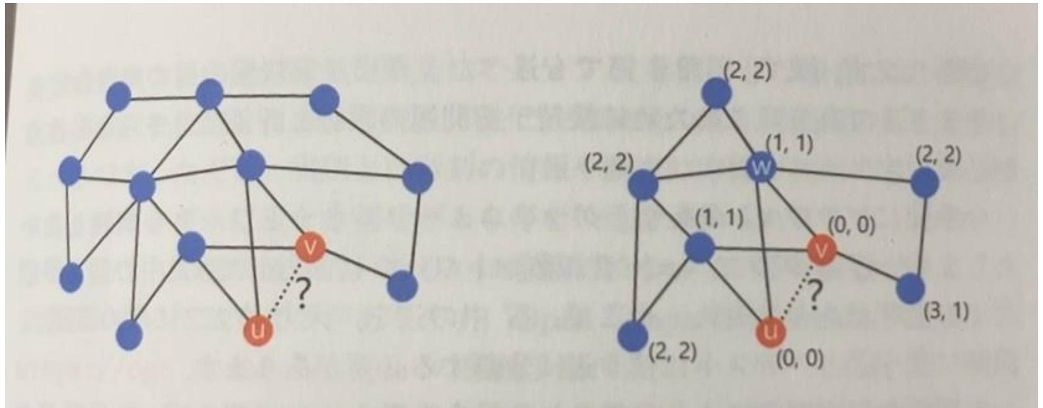


図 4.2 接続予測をグラフ分類に帰着する方法．左：入力グラフ G．頂点 u と頂点 v の間に辺が存在するかどうかを予測したいとする．右：切り出したグラフ $G(\mathcal{N}_2(\{u, v\}))$．このグラフを正例・負例に分類する．各頂点に付記しているのは構造に基づいたグラフ特徴量である．第 1 成分は u からのグラフ上の距離，第 2 成分は v からのグラフ上の距離を表す．この特徴量は予測対象の頂点対 $\{u, v\}$ に依存する．

4.2 接続予測 -4

頂点予測問題における注意点

- データリークが起こりやすい
  - ■ 時系列のデータ
  - ■ モデルに入力するグラフと、学習に使用する辺の混同
- 辺（グラフ）を観測できないことがある
  - ■ 例えば、学習予定のグラフを訓練、評価、テストに分割し、十分に予測可能かをみる（予測できない場合、データに不具合がある）

STEP 3

実際のコードを見る