5L-03

確率分布比推定を用いたロバストなグラフ埋め込み

薩田　凱斗 †　　佐々木 博昭 ‡
公立はこだて未来大学 †　公立はこだて未来大学 ‡

1　背景

グラフデータとは，グラフの各頂点に与えられたデータベクトルと頂点間の辺の重みから構成されるデータであり，近年，様々な場面で取得されている．例えば，頂点がユーザー，辺がユーザー間の交友関係等を表すソーシャルネットワーク [1] や，頂点が論文，辺が引用関係を表す論文引用ネットワーク [2] 等が挙げられる．

グラフデータを扱う上で重要なタスクは，グラフ埋め込みと呼ばれるデータベクトルの特徴写像を学習することである．この特徴写像を学習することで，既存のデータ解析手法をグラフデータへ適用することが可能となる．特徴写像を学習するための 1 つのアプローチは，最尤推定法に基づくアプローチであろう [3]．しかし，最尤推定法では条件付き確率分布のモデル化が必要であり，確率分布のモデルが真の確率分布と大きく異なる場合，良い特徴写像が学習されない可能性がある．それに加え，最尤推定法はデータの外れ値の影響を強く受ける傾向にある．これらの問題に対して，[4] は，β-クロスエントロピーを用いた条件付き平均推定を通じて，β-グラフ埋め込み（以下，β-GE）と呼ばれる手法を提案した．β-クロスエントロピーにより，β-GE では外れ値にロバストな特徴写像の学習が可能となり，数値実験により，外れ値に対するロバスト性が示されている．その一方で，条件付き確率分布と比較すると，条件付き平均は辺の重みとデータベクトル間の限定的な依存関係しか捉えることができない．

そこで，本研究では，外れ値にロバストかつより一般性の高い依存関係を捉えるグラフ埋め込み手法を提案する．辺の重みとデータベクトル間の依存関係を捉えるために，条件付き確率分布と周辺確率分布の比の推定を通じて特徴写像を学習する．これにより，β-GE と比較して，より一般的な依存関係を捉えることができ，より広範囲なデータに対して有効な手法となることが期待される．加えて，外れ値に対してロバストな手法を提案するために，γ-クロスエントロピー [5] を用いた学習を行う．人工データに基づく数値実験により，提案法が外れ値に対してロバストであり，既存手法より優れた性能をもつことを示す．

2　確率分布比推定によるロバストなグラフ埋め込み

2.1　問題設定

本研究で扱う n 個の頂点と頂点間の辺からなる無向グラフ上のグラフデータについて述べる．頂点 i に d 次元のデータベクトル $\boldsymbol{x}_i$, 頂点 i と頂点 j 間には辺の重み $w_{ij}(\geq 0)$ が与えられ，これら $w_{ij}, \boldsymbol{x}_i, \boldsymbol{x}_j$ は同時確率分布 $p(w_{ij}, \boldsymbol{x}_i, \boldsymbol{x}_j)$ に従い，生成されるとする．このとき，グラフ埋め込みの目的は，$\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{w_{ij}\}_{i,j=1}^n$ からデータベクトル $\boldsymbol{x}$ の特徴写像 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ を学習することである．ここで，$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ は d 次元から K 次元への写像である．

2.2　提案手法

本研究では，以下の条件付き確率分布 $p(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ と周辺確率分布 $p(w_{ij})$ の比（以下，分布比）を特徴写像 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ を用いてモデル化することでパラメータ $\boldsymbol{\theta}$ の推定を行う．

$$\log \frac{p(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)}{p(w_{ij})}.$$

Robust graph embedding with distribution ratio estimation
† Kaito Satta, FutureUniversity Hakodate
‡ Hiroaki Sasaki, Future University Hakodate

条件付き平均 $E(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ とは異なり，本研究における分布比推定では，条件付き確率分布 $p(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ をより直接的に扱う．したがって，$E(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ が定数となるような確率分布に対しても提案手法は有効なグラフ埋め込み手法となることが期待される．

次に，外れ値に対して頑強な手法を提案するために，γ-クロスエントロピー [5] を用いる．具体的には，[6] で提案された手法をグラフデータへ応用し，次の目的関数を最小化することで，パラメータ推定を行う．

$$\widehat{J}_\gamma(\boldsymbol{\theta}) := -\frac{1}{\gamma}\log\left[\frac{1}{n^2}\sum_{ij}\left(\frac{e^{(\gamma+1)w_{ij}\langle f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_j)\rangle}}{1+e^{(\gamma+1)w_{ij}\langle f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_j)\rangle}}\right)^{\frac{\gamma}{\gamma+1}} + \frac{1}{n^2}\sum_{ij}\left(\frac{1}{1+e^{(\gamma+1)w^*_{ij}\langle f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_j)\rangle}}\right)^{\frac{\gamma}{\gamma+1}}\right] \quad (1)$$

式 (1) の w^*_{ij} は w_{ij} を i, j に関してランダムにシャフルした辺の重み，$\langle\cdot,\cdot\rangle$ は内積を表す．γ は正のパラメータであり，γ が大きくなるにつれて，推定がロバストになる傾向にある．目的関数 $\widehat{J}_\gamma(\boldsymbol{\theta})$ を最小化する $\boldsymbol{\theta}$ を $\widehat{\boldsymbol{\theta}}$ とすると，$\boldsymbol{f}_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})$ が提案法のグラフ埋め込み（特徴ベクトル）である．

3 数値実験

本章では，提案法を頂点クラスタリングへ応用し，その有効性を人工データを用いた数値実験により確認する．頂点数 $n = 200$ とし，頂点 i のデータベクトル $\boldsymbol{x}_i$ を 4 つの等方的なガウス分布から構成される混合ガウス分布から生成した．したがって，4 つのガウス分布の平均がクラスタ中心となる．

次に，生成されたデータベクトル $\boldsymbol{x}_i, \boldsymbol{x}_j$ に基づき，辺の重み w_{ij} を次のように 2 通りで生成した．**（ベルヌーイ）**確率 p のベルヌーイ分布を $B(p)$ とすると，頂点 i と j が同じクラスタの場合，$B(0.05)$ から辺の重み w_{ij} を生成，一方，異なるクラスタの場合，$B(0.03)$ から辺の重みを生成した．したがって，$B(0.03)$ から生成された w_{ij} が外れ値と解釈できる．**（切断正規）**$-0.5 \leq \tilde{w}_{ij} \leq 2.5$ 上の切断正規分布から $\tilde{w}_{ij}$ を生成し，$\tilde{w}_{ij}$ を $0, 1, 2$ へと離散化した値を w_{ij} の値とした．その切断正規分布の中心は定数 1，その広がりは $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ の関数として表現されている．したがって，このデータの条件付き平均 $E(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ は近似的に定数とみなすことができる．生成したグラフデータに対して，提案法を適用し，特徴ベクトルを学習した．最後に，学習した特徴ベクトルに対して，k 平均クラスタリングを適用し，そのクラスリング結果を調整ランド指数 [7] により評価した．

表 1　調整ランド指数 [7] の平均とその標準誤差

	ベルヌーイ	切断正規
ML-GE	0.66 ± 0.03	0.01 ± 0.00
β-GE	$\mathbf{0.81 \pm 0.03}$	0.27 ± 0.06
提案法	0.80 ± 0.03	$\mathbf{0.83 \pm 0.03}$

比較手法として，β-GE に加え，最尤推定法によるグラフ埋め込み（ML-GE）を用いたクラスタリング結果を表 1 に示す．この結果から，提案法が β-GE と同程度にロバストであり，条件付き期待値が定数となるような確率分布では比較手法よりも有効な手法であることが分かる．

4 結論

本研究では，外れ値にロバストなグラフ埋め込み法を分布比推定に基づき提案した．数値実験により，提案法が外れ値にロバストであり，条件付き平均推定に基づく手法よりも有効であることを確認した．

参考文献

[1] L. David and K. Jon, JASIST, 2007.
[2] P. Sen et al., AI Mag., 2008.
[3] A. Okuno et al., Proc. of ICML, 2018.
[4] A. Okuno et al., Proc. of AISTATS, 2019.
[5] H. Fujisawa, and S. Eguchi, JMVA, 2008.
[6] H. Sasaki and T. Takenouchi, arXiv, 2020.
[7] H. Lawrence and P. Arabie, J. Classif., 1985.