DEIM2022 G21-5 (day2 p36)

階層型クラスタ構造を活用する部分グラフベースのリンク予測

中西　宏和† 　山口　寛人† 　前川　政司† 　佐々木　勇和† 　鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5

E-mail: †{nakanishi.hirokazu,yamaguchi.hiroto,maekawa.seiji,sasaki,onizuka}@ist.osaka-u.ac.jp

あらまし　グラフにおけるリンク予測は， SNS の友人推薦や将来の消費者の購買予測など，様々なアプリケーションに応用されている．既存手法としては，部分グラフベースのリンク予測手法が現在の主流である．しかし，部分グラフを用いたリンク予測では局所的なグラフ構造しか活用できないという問題がある．そのため，それら手法に対し大域的なグラフ構造を表す特徴量を付加することで，リンク予測の精度向上を目指す． 提案手法では，階層型クラスタリングを全体のグラフに適用し，そのクラスタリング結果から作成した特徴量を部分グラフベースの手法（SEAL）に追加することで，大域的なグラフ構造を活用したリンク予測を実現する．本稿では 2 つの特徴量を提案する．一つ目は，部分グラフ中の各ノードが部分グラフ中の各ノードとリンク予測対象の各ノードがともに属する最小クラスタを表す「最小クラスタサイズ」である．二つ目は部分グラフ中の各ノードがどのクラスタに属すかを表す「クラスタベクトル」である．実験では，クラスタベクトルを活用することで，5 つのデータセットのうち 4 つのデータセットで既存手法を上回る精度を達成した．

キーワード　リンク予測，グラフクラスタリング，ノードエンベディング

1　はじめに

グラフとは，ノード（頂点）とリンク（辺）によって表現されるデータ構造であり，物事の関係を表現することができる．グラフの例としては，ソーシャルネットワークや神経ネットワーク，論文の引用グラフなどがある．こうしたグラフに対して，その中にあるコミュニティを抽出するクラスタリング，各ノードのクラスの分類を行うノード分類，2 つのノード間にリンクが存在するかを予測するリンク予測など，多岐に渡った解析が行われている [1–3]．これらは様々なアプリケーションに応用され，研究が盛んに行われている．本論文で取り扱うのはこのうちのリンク予測である．リンク予測の応用例は様々であり，SNS における友人推薦や将来の消費者の購買予測，代謝ネットワーク再構築などが挙げられる．

リンク予測の初期の手法には，似たノード間にリンクが張られるという仮定に基づいたヒューリスティック [4–7] や，これらの改良手法として最近は GNN （グラフニューラルネットワーク）に基づいた手法 [3, 8] が提案されている．後者の中でも，部分グラフを用いた手法 [3, 9, 10] は予測対象のノードペアの近隣情報を有効に活用することによって，高い精度が得られている．そのため，これら部分グラフを用いた手法に関して現在幅広く研究が行われており， SEAL [3] はその代表例といえる．SEAL では，予測対象のノードペアからのホップ数によって決定される特徴量である「構造ノードラベル」が提案されている．これを用いて GNN の学習を行うことにより，リンク予測対象のノードペアの近傍のグラフ構造を活用することができ，リンク予測における精度向上に大いに貢献していると知られている．

しかし，このような部分グラフを用いる手法では，部分グラフからリンクの有無を予測するため，局所的なグラフ構造を積極的に活用している一方で，グラフにおける重要な性質の一つであるコミュニティ性を活用できておらず，大域的なグラフ構造の情報が活用されていないと考えられる．

本論文では，部分グラフを用いる手法に対して，大域的なグラフ構造を表現する特徴量を明示的に付加することで，リンク予測の精度向上を目指す．クラスタリングにおける，同一クラスタ内のリンクは密になり，クラスタ間のリンクは疎になるという性質を活用して，提案手法における大域的なグラフ構造として，そのクラスタリング結果を反映する特徴量を作成し，それを活用することでリンク予測の精度向上を図る．また，クラスタリングの中でも，階層型クラスタリングを用いれば，様々な粒度のクラスタ情報を抽出することができるため，本手法ではこの階層型クラスタリングを用いる．得られたクラスタリング結果からは，階層型クラスタ構造を反映する二つの特徴量を提案する．一つ目は，部分グラフ中の各ノードが部分グラフ中の各ノードとリンク予測対象の各ノードがともに属する最小クラスタを表す「最小クラスタサイズ」である．この特徴量は，ホップ数では離れているが，同クラスタに属するノード間のリンク発生確率を高く見積もることを狙いとしている．二つ目は，部分グラフ中の各ノードがどのクラスタに属すかを表す「クラスタベクトル」である．この特徴量は，クラスタリング結果をそのまま反映し，理解しやすいものとしている．これらの階層型クラスタ構造を表現する特徴量（以後これらをクラスタ特徴量と総称する）を，それぞれ SEAL に組み込んで学習を行うことで提案手法を実装する．

実験では，まず，提案手法のリンク予測精度を既存手法である SEAL と比較する実験を行った．これにより，二つ目に提案した特徴量であるクラスタベクトルを活用することで，既存手法に比べてリンク予測精度が向上することを確認した．また，クラスタベクトルの算出時に指定するパラメータを変化させて

リンク予測精度を測定するパラメータセンシティビティの実験を行った．

本稿の構成は以下の通りである． 2 章で事前準備を行い， 3 章で提案手法について詳細に述べた後， 4 章で実験結果を示し，考察を行う．また， 5 章では既存のリンク予測に用いられる関連研究を紹介する．最後に， 6 章で結論を述べる．

2 事 前 準 備

2.1 グ ラ フ

本論文で用いるグラフは重み無しの無向グラフであり，以後これをグラフと呼ぶことにする．グラフ G はノード数を n とすると，ノード集合 $V = \{v_1, v_2, ..., v_n\}$ とリンク集合 $E = \{v_i, v_j\} \subseteq [V] \times [V]$ により，$G = (V, E)$ と表される．グラフの隣接行列 $A \in \mathbb{R}^{n \times n}$ とは，ノード v_i とノード v_j の間にリンクが存在する場合は $A_{i,j} = 1$，リンクが存在しない場合には $A_{i,j} = 0$ と定義される行列である．

2.2 グラフの統計量

グラフ全体の統計量として，平均次数と平均クラスタ係数（ACC） [11] がある．まず，次数とは，ノードにつながっているリンクの数を表す．ノード v_i の次数は，$deg(v_i)$ と表記される．平均次数は，グラフ全体における各ノードの次数の平均値であり，値が大きいほどノード数に対してリンク数が多いグラフであるといえる．平均次数は，ノード数を n とすると，$\frac{1}{n}\sum_i deg(v_i)$ と表される．次に、クラスタ係数とは，一つのノードに対して定義できる値であり，隣接ノードにリンクが張られている割合を表す．ノード v_i のクラスタ係数 C_i は，ノード i から張られているリンクの総数を E_i とすると，

$$C_i = \frac{2E_i}{deg(v_i)\{deg(v_i) - 1\}} \tag{1}$$

と表される．各ノードのその平均値を計算したものが平均クラスタ係数であり，$\frac{1}{n}\sum_i C_i$ と表される．これは，ネットワークがどの程度密につながっているかを表し，値が大きいほど密なグラフだといえる．

2.3 リンク予測

リンク予測の問題定義を行う．本研究においては，観測されていない潜在的なリンク（missing link）を予測することをリンク予測の目的とする．n 個のノードを持つ観測されたグラフ $G_0 = (V, E_0)$ を考える．E_0 は観測されたリンクの集合とする．観測されたリンクの集合 E_0 は、本来のグラフを $G = (V, E)$ としたとき，$E_0 \subseteq E$ である．リンク予測の目的は，E に含まれる真のリンクと E に含まれない偽のリンクを含む候補セット E_C の中から，真のリンクを推論することである．

2.4 SEAL

SEAL は，予測対象のノードペアの近傍の部分グラフの抽出を行い，グラフの局所情報を用いてリンクの有無を予測する手法であり，最先端のリンク予測手法である．局所的なグラフ構造を反映する構造ノードラベルを活用することにより，非常に高い精度を実現している．

SEAL の学習には 三つのフェーズある． 第一フェーズでは enclosing subgraph の構築，第二フェーズではノード情報行列の作成．第三フェーズでは グラフニューラルネットワーク（GNN） の学習を行う．

第一フェーズにおいて，ノード v_x とノード v_y の最短距離を $\Delta(x, y)$ とすると，ノード v_i と v_j に対する h ホップまでの enclosing subgraph を構成するノードの集合 $V^h_{\{i,j\}}$ は以下のように表せる．

$$V^h_{\{i,j\}} = \{x \in V | \Delta(x, i) \leqq h \, or \, \Delta(x, j) \leqq h\} \tag{2}$$

これより，enclosing subgraph のリンクの集合 $E^h_{\{i,j\}}$ は以下のように表せる．

$$E^h_{\{i,j\}} = \{\{x, y\} \in [V] \times [V] | x, y \in V^h_{\{i,j\}}\} \tag{3}$$

これらより，enclosing subgraph $G^h_{\{i,j\}}$ は， $G^h_{\{i,j\}} = (V^h_{\{i,j\}}, E^h_{\{i,j\}})$ と表される．なお，SEAL では，基本的に $h = 2$ が最も適切なパラメータとされている．第二フェーズでは，各 enclosing subgraph から導出される構造ノードラベルと，ノード属性，ノードエンベディングを各 enclosing subgraph 中のそれぞれのノードごとに結合し，ノード情報行列 X_{info} を作成する．構造ノードラベルとは，各ノードに割り当てられる整数のラベルであり， 対象リンクを構成するノード v_i とノード v_j との最短距離により決定される特徴量である．enclosing subgraph の中心がノード v_i と v_j のとき，その enclosing subgraph 中におけるノード $v_k \in V^h_{\{i,j\}}$ の構造ノードラベル $f_l(k)$ は以下の式で表される．

$$\begin{aligned} f_l(k) =& 1 + min(\Delta_i, \Delta_j) \\ &+ \frac{\Delta}{2}[\frac{\Delta}{2} + (\Delta \; mod \; 2) - 1] \end{aligned} \tag{4}$$

なお，$\Delta_i = \Delta(v_i, v_k)$，$\Delta_j = \Delta(v_j, v_k)$，$\Delta = \Delta_i + \Delta_j$ である．ただし，$k = i, j$ のときは $f_l(k) = 0$ である．また，二ノード間に経路が存在しない場合も $f_l(k) = 0$ である．なお，ノード属性とノードエンベディングはノード v_k に対応する行を抽出してきたものを用いる．enclosing subgraph のノード数が n_{es}，構造ノードラベルの次元数が 1，ノード属性の次元数が d_a，ノードエンベディングの次元数が d_e のとき，最終的なノード情報行列 X は，$X \in R^{n_{es} \times d_{es}} (d_{es} = 1 + d_a + d_e)$ となる．第三フェーズでは，第二フェーズで作成したノード情報行列 X を持つ enclosing subgraph を DGCNN [12] などのグラフニューラルネットワーク（GNN , グラフ深層学習）に入力し，リンクの有無を予測する．

SEAL の特徴としては，グラフの構造だけでなく，ノード属性や node2vec [13] 等によるノードエンベディングを各ノードの情報として組み込むことが挙げられる．また，リンクの存在しないノードペアをネガティブリンクとして学習に用いることで精度を向上させていることが判っている．これらの点におい

て， SEAL の先駆けとなる手法である WLNM [9] よりも改善された手法であるといえる．

2.5 Louvain 法

Louvain 法は大規模なネットワークに対しても短時間で実行できる階層型グラフクラスタリング手法の一つであり，様々な粒度のコミュニティを検出することができる．Louvain 法では，Modularity [14] を目的関数に用いる．Modularity とは，クラスタ間のリンクに対する，クラスタ内のリンクの密度を測定する指標であり，−1 から 1 のスカラー値となる．以下にその定義式を示す．

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \tag{5}$$

c_i はノード v_i が属するクラスタである．k_i はノード v_i に結合しているリンクの重みの合計であり，$k_i = \Sigma_j A_{ij}$ と表される．$\delta(u, v)$ は $u = v$ のとき 1 でそれ以外は 0 となるクロネッカーのデルタである．m は全てのリンクの重みの合計を総合したものであり，$m = \frac{1}{2}\Sigma_{ij} A_{ij}$ と表される．Modularity は，グラフクラスタリングの性能評価に用いられる他，本手法のように Modularity 自体を目的関数として用いる場合もある．

この手法のアルゴリズムは二つのフェーズに分かれており，繰り返し実行される．第一フェーズでは，各ノードに個別のクラスタを与える．すなわち，最初はノード数を n とすると，それと同数の n 個のクラスタが存在することとなる．次に，ノード v_i に隣接するノード $\{v_j \in V | A_{ij} = 1\}$ のクラスタすべてについて，ノード v_i を加えた場合に生じる Modularity の利得 ΔQ を算出する．この利得が最大となるクラスタにノード v_i を加える．このプロセスをすべてのノードに対して，Modularity が改善されなくなるまで繰り返して第一フェーズは終了する．

第二フェーズでは，第一フェーズで見つけたクラスタをノードとする新しいグラフを構築する．新しいノード間のリンクの重みは対応する 2 つのクラスタ間のリンク数の総和となる．同じクラスタ内のノード間のリンクは自己ループとなる．第二フェーズで得られた新しいグラフを第一フェーズから再適用することで反復することが可能となる．

本実験で用いる Louvain 法によるクラスタリング結果は，最終的なクラスタだけでなく，デンドログラムを解析することにより，中間生成されるクラスタも含んでいる．そのため，様々な粒度のコミュニティの情報をクラスタリング結果として得られる．このようなクラスタの情報を使うことで，様々な粒度のコミュニティの情報をクラスタリング結果として得ることができると考えられる．なお，後の提案手法の実装の都合で，最終的に生成されるクラスタを全て集約した，つまり，全てのノードを含むクラスタを最上位のクラスタとなるように工夫を施すこととする．

3 提 案 手 法

3.1 提案手法モデル

提案手法の概要図を図 1 に示す．本稿では，部分グラフベースの手法として代表的な手法である SEAL をベースに提案手法を構築する．グラフを入力として enclosing subgraph の抽出を行う一方で，提案手法ではクラスタ特徴量の算出も行う．大域的なグラフ構造を獲得するために，全体のグラフに対して階層型クラスタリング手法である Louvain 法を適用する．クラスタリングには，同一クラスタ内のリンクは密になり，クラスタ間のリンクは疎になるという性質がある．そのため，クラスタリングの結果を用いることで，リンク予測の精度向上につながると考えられる．また， Louvain 法は階層型クラスタリングでもあり，様々な粒度のクラスタ構造を獲得することができるため，大域的なグラフ構造を反映することができる．得られたクラスタリング結果に基づき，大域的なグラフ構造を反映する二つのクラスタ特徴量として,「最小クラスタサイズ」と「クラスタベクトル」を提案する．最小クラスタサイズは，部分グラフ中の各ノードがリンク予測対象のノードペアからどれだけ近いクラスタに属しているかを表す．一方，クラスタベクトルは，部分グラフ中の各ノードがどのクラスタに所属しているかを表す．上記の各特徴量を SEAL におけるノード情報行列 X に対して結合することで，大域的なグラフ構造情報を付与する．以下，最小クラスタサイズを特徴量として付与した手法を提案手法 MCS ，クラスタベクトルを付与した手法を提案手法 CV と呼ぶ．このクラスタ特徴量を付与したノード情報行列を持った enclosing subgraph を，グラフニューラルネットワークである DGCNN に入力し，最終的なリンク予測結果を得る． DGCNN は，入力されたグラフのラベルを当てるというグラフレベルのクラス分類（グラフ分類問題）を行うアルゴリズムであるが， SEAL では，リンクの有無をラベルと見立ててることで，リンク予測問題に帰着している．

3.2 最小クラスタサイズ

最小クラスタサイズは，enclosing subgraph 中の各ノードがリンク予測対象のノードペアからどれだけ近いクラスタに属しているかを表す特徴量である．最小クラスタサイズの概要を図 2a に示す．最小クラスタサイズでは，大域的なグラフ構造を捉えるために，クラスタをベースとした二つのノード間の近さを表現するように設計している．これにより，ホップ数では離れているが，同クラスタに属する二ノード間のリンク発生確率を高く見積もることができると考えられる．近さを定義するために，二つのノードがともに属する最小クラスタのサイズを活用する．以上より， SEAL における構造ノードラベルでは捉えることができない，より大域的なグラフ構造を特徴量に埋め込むことが可能となり，リンク予測の精度向上に貢献すると考えられる．

ノード v_i とノード v_j がリンク予測対象のノードペアである

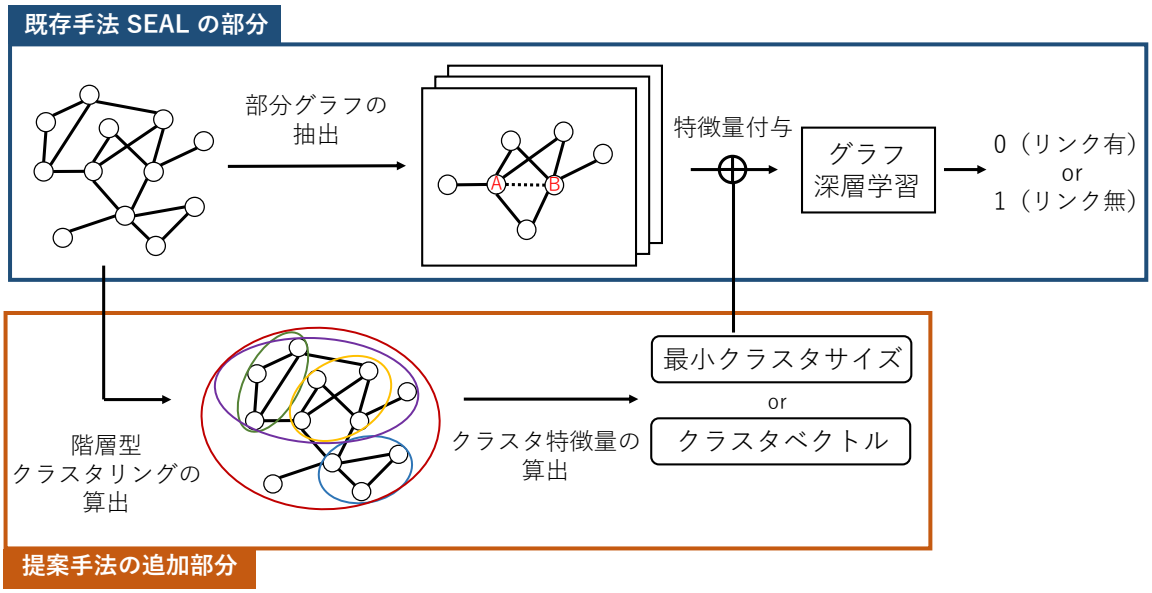


図 1: 提案手法のモデル概要図．提案手法は，グラフとリンク予測対象のノードペアを入力とし，リンク予測結果を出力する．まず，入力されたグラフに基づき，リンク予測対象のノードに対応する enclosing subgraph の抽出と階層型クラスタリングが行われる． なお，図中の赤字のノード v_A ，v_B はリンク予測対象のノードペアを表し， enclosing subgraph は簡単のため 1 ホップとして図示している．また，色のついた楕円で囲まれたノード群は対応した色のクラスタに属していることを示している．得られたクラスタリング結果からクラスタ特徴量を算出し，それを enclosing subgraph のノード情報行列に連結する．最後に，グラフ深層学習によりリンク予測の学習及び推論を行う．

とき，enclosing subgraph 内のノード $v_k \in V^k_{\{i,j\}}$ の最小クラスタサイズ $\boldsymbol{x}_k$ は以下のように定義される．

$$\boldsymbol{x}_k = [\min_{c_l \in C}\{size(c_l)|v_i \in c_l, v_k \in c_l\}, \\ \min_{c_l \in C}\{size(c_l)|v_j \in c_l, v_k \in c_l\}] \tag{6}$$

クラスタの集合を C としたとき，各クラスタを $c_l \in C$ と表記する．$size(\cdot)$ はクラスタのサイズを表す．クラスタはノード $v_i \in V$ を要素とする集合とする．最小クラスタサイズは 2 次元のベクトルとなる．一つ目の次元はノード v_i とノード v_k がともに属するクラスタの最小サイズ， 2 つ目の次元はノード v_j とノード v_k がともに属するクラスタの最小サイズとなっている．$\boldsymbol{x}_k$ を enclosing subgraph のノード数 n_{es} だけ並べて行列とした，$X_{mcs} \in \mathbb{R}^{n_{es}\times 2}$ が enclosing subgraph の各ノードの最小クラスタサイズを並べたものとなる．これを各 enclosing subgraph ごとにノード情報行列に連結する．なお，連結する node2vec の各要素の値域が $[-1, 1]$ であるのに対して，最小クラスタサイズをそのまま適用すると正の整数値である．これでは， node2vec の値に比べて，クラスタサイズの値が非常に大きくなりクラスタサイズの影響も非常に大きくなってしまう．そのため，実際には最も大きなクラスタサイズで除算して，クラスタサイズの各要素の値域が $[0, 1]$ となるように正規化する．なお， Louvain 法の実装の工夫により，最も大きなクラスタはすべてのノードを含むクラスタとなるようにしてある．

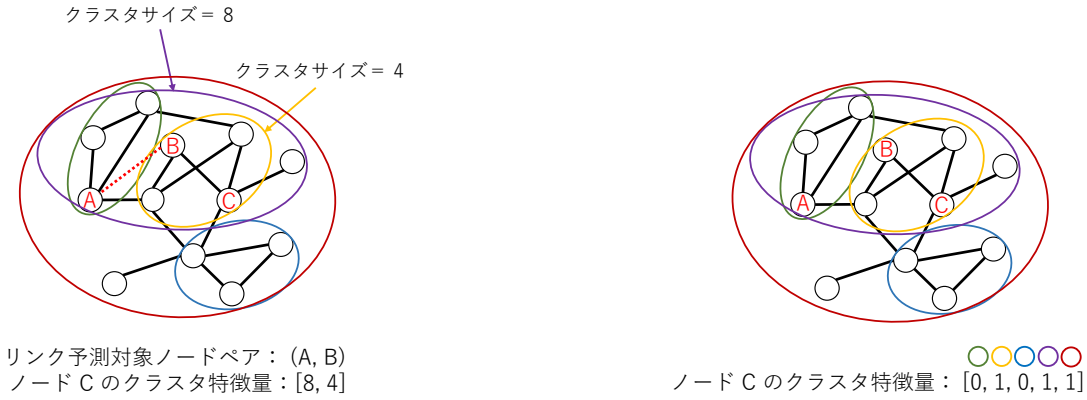
最小クラスタサイズは， enclosing subgraph のリンク予測対象のノードペアが異なるならば，同一ノードの最小クラスタサイズであってもその値は異なる．つまり，リンク予測対象のノードペアに依存する特徴量となっている．その反面，クラスタリング結果を単純に反映したものではないため，大域的なグラフ構造の情報をすべて網羅できているとはいえず，後述のクラスタベクトルよりも情報量が少ない可能性があると考えられる．

3.3 クラスタベクトル

クラスタベクトルは，enclosing subgraph 中の各ノードがどのクラスタに属すかを表現したものである．クラスタベクトルの概要を図 2b に示す．設計指針としては，ノードが属する局所的なクラスタから大域的なクラスタまでの階層構造を表現することを目的としている

異なる階層を含めたクラスタの総数を q とすると，ノード v_i のクラスタベクトル $\boldsymbol{x}_i$ は，次元数が q で，$v_i \in C_l$ のとき，$\boldsymbol{x}_k$ の l 番目の要素が 1，そうでなければ 0 の値をとる．これを enclosing subgraph のノード数 n_{es} だけ並べて行列とした，$X_{cv} \in R^{n_{es}\times q}$ が enclosing subgraph における各ノードのクラスタベクトルを全て並べたものとなる．これを各 enclosing subgraph ごとにノード情報行列に連結する．

この特徴量は，クラスタリングの結果を単純にベクトルに変換しているため，実装が容易であり，理解することも簡単である．また，最小クラスタサイズとは対照的に， enclosing subgraph のノードペアによって変化することはなく，各ノードで固有の値を持つ．そのため， enclosing subgraph ごとに再計算する必要はなく，一度算出すればそれを再利用することができる．ただし，巨大なネットワークの場合，次元数が非常に大きくなり，過学習が起きやすくなる．よって，本実験ではクラスタベクトルが一定の次元数を超える場合，下位階層のクラスタに対応する次元から一定数だけを用いる．これは，下位階層のクラスタを優先的に用いる方が良いことが補足実験によ



(a) 最小クラスタサイズ　　**(b)** クラスタベクトル

図 2: 最小クラスタサイズとクラスタベクトルの概要図．図では，ノード v_C のクラスタ特徴量をそれぞれ算出している．実際には各 enclosing subgraph 中の全てのノードに対して算出される．最小クラスタサイズについては，リンク予測対象のノードペアがノード v_A，v_B の場合のノード v_C の最小クラスタサイズを表す．このとき，一つ目の次元は，ノード v_A とノード v_C がともに属する最小のクラスタのサイズの 8 であり，二つ目の次元は，ノード v_A とノード v_C がともに属する最小のクラスタのサイズの 4 である．ただし，実際には最大のクラスタサイズにより除算が行われ，値域は $[0, 1]$ となる．クラスタベクトルについては，各次元はその上の円の色に対応するクラスタに対応している．ノード v_C は黄と赤のクラスタに属しているため，その次元の値は 1 で，それ以外の次元は 0 となっている．

り分かっているためである．このクラスタベクトルの上限の次元数のことを本論文では上限次元数と定義し，パラメータとして扱う．

4 実　　験

本稿では，以下の二つの実験を行う．はじめに，提案手法のリンク予測精度と既存手法の精度を比較する．これにより，大域的なグラフ構造を活用する提案手法の有効性を確認する．次に，提案特徴量クラスタベクトルのパラメータである上限次元数を変化させてリンク予測精度を測定し，有効なパラメータについて調べる．これにより，上限次元数がリンク予測精度に与える影響について確認する．

4.1 データセット

本実験で用いるデータセットについての説明を以下に示す．

- **NS**：ネットワーク科学研究者の共著グラフ．
- **USAir**：米国の航空会社のグラフ．
- **Power**：米国西部の電力網ネットワーク．
- **Celegans**：線虫の生物学的ネットワーク．
- **Router**：ルータレベルのインターネット．

各データセットの統計量を表 1 に示す．なお，これらデータセットはノード属性を持たないネットワークである．

4.2 実 験 設 定

実験設定は基本的には SEAL の論文におけるものと同一である．データセットは，train データが 80%，validation データが 10%，test データが 10% となるように分割する．学習は epoch が 100 に達するまで実行を行い，各 epoch について validation データにおける loss が最小となった epoch のモデルを選択する．そのモデルに対して test データでリンク予測を行った際の結果を実験結果とする．

ノード情報行列に付与するノードエンベディングには，SEAL の実験に倣って node2vec を用いている．グラフニューラルネットワークには DGCNN を用い，sortpooling のパラメータは 0.6 である．また，リンク予測対象のノードペアのうち，リンクが存在するノードペアのデータを正例，リンクが存在しないノードペアのデータを負例とすると，学習においてリークを防ぐため，test データと validation データの正例のリンクは削除し，負例に関してはなにも施さない．

4.3 比 較 実 験

既存手法と提案手法の精度の比較を行う．データセットは五つ全てのデータセットを用い，評価指標には，Accuracy と AUC [15] を用いる．乱数シードを変更することで，それぞれ 10 回試行し，その平均の値を各結果として用いる．また，その際の標準偏差を ± で表している．比較手法には提案手法の原型となる SEAL を用いる．なお，クラスタベクトルにおけるパラメータである上限次元数は，NS では 32，Power では 128，Router では 16 としている．既存手法 SEAL と提案手法 CV，提案手法 MCS の精度を比較した結果を表 2 と表 3 に示す．

提案手法 CV では，Celegans の AUC を除いたすべての結果において，既存手法に比べて提案手法の精度の向上が確認できた．このことから，SEAL にクラスタベクトルを追加して学習することの有効性が確認できる．つまり，大域的なグラフ構造を表す特徴量としてクラスタベクトルが適切な特徴量であると考えられる．精度向上が実現した理由としては，構造ノードラベルでは捉えることができない，グラフの重要な性質の一つであるコミュニティ性をクラスタベクトルが情報として含んでいるためだと考えられる．しかし，標準偏差を考慮すると，NS と Router 以外の結果はあまり有意なものではないと考えられる．これは，SEAL の精度自体が非常に高水準であるということが考えられる．その他にも，USAir と Celegans 以外

データセット名	ノード数	リンク数	平均次数	ACC	Modularity	クラスタ数
NS	1589	2742	3.45	0.638	0.9597	1045
USAir	332	2126	12.81	0.625	0.3508	121
Power	4941	6594	2.67	0.080	0.9359	2426
Celegans	297	2148	14.46	0.292	0.3876	106
Router	5022	6258	2.49	0.012	0.9038	2384

表 1: データセットの詳細．ノード数とエッジ数，平均次数， ACC ， Modularity の値をそれぞれ示す．なお，Modularity の値は Louvain 法を適用した際のクラスタリング結果に対して算出されたものである．

データセット	SEAL	提案手法 MCS	提案手法 CV
NS	90.84 ± 1.30	91.30 ± 0.93	$\mathbf{91.86} \pm 0.74$
USAir	91.91 ± 0.86	90.15 ± 0.80	$\mathbf{92.47} \pm 0.60$
Power	74.76 ± 0.52	72.10 ± 2.47	$\mathbf{74.89} \pm 0.49$
Celegans	79.84 ± 1.43	79.14 ± 0.92	$\mathbf{79.96} \pm 1.00$
Router	83.43 ± 0.45	70.60 ± 0.88	$\mathbf{84.88} \pm 0.52$

表 2: 比較実験の Accuracy の結果．各データセットにおいて精度が最も良いものは太字で表している．

データセット	SEAL	提案手法 MCS	提案手法 CV
NS	96.61 ± 0.50	96.34 ± 0.41	$\mathbf{97.23} \pm 0.31$
USAir	97.29 ± 0.39	96.09 ± 0.33	$\mathbf{97.48} \pm 0.34$
Power	81.39 ± 0.36	79.02 ± 0.77	$\mathbf{81.82} \pm 0.45$
Celegans	$\mathbf{88.07} \pm 0.90$	86.85 ± 0.35	87.57 ± 0.71
Router	92.37 ± 0.44	75.87 ± 1.16	$\mathbf{93.65} \pm 0.30$

表 3: 比較実験の AUC の結果．各データセットにおいて精度が最も良いものは太字で表している．

のデータセットでは，上限次元数による次元数削減により，上位階層のクラスタ情報を活用することができておらず，精度が良くならない可能性が考えられる．そのため，PCA [16] などを用いて次元削減を行うことで次元数を抑えつつ情報の密度を高めることで，精度が改善される可能性があると考えられる．

一方，提案手法 MCS では， NS データセットにおいては既存手法に比べて精度の向上が確認されたが，そのほかのデータセットでは精度が低下した．このことから，最小クラスタサイズは大域的なグラフ構造を表現する特徴量として不適切であることが考えられる．この理由として，学習においてリークを防ぐために test データのリンク予測対象のノードペアのうち正例のリンクを削除していることが，悪影響を及ぼす原因であると考えられる．train データにおける正例については，予測対象のノードペアのリンクを削除しないため，そのノードペアが階層的に近いクラスタに属しやすい．一方， test データにおける正例は，予測対象のノードペアのリンクを削除するため，正例に比べてそのノードペアが階層的に遠いクラスタに属しやすい．そのため， test データの正例は， train データに比べて最小クラスタサイズの値が大きくなってしまうと考えられる．これでは， train データと test データで正例の特徴量の傾向が異なってしまい，学習はうまく行われたとしても，推論がうまくいかないと考えられる．

4.4 パラメータセンシティビティ

提案手法 CV にはクラスタベクトルのパラメータとして上限次元数がある．このパラメータがリンク予測結果に与える影響を調べるため，クラスタリングにより得られるクラスタ数が大きいデータセットである NS ， Power ， Router について，パラメータを変化させた場合の精度の変化を確認する実験を行った．上限次元数は，16，32，64，96，128，160，192，256，384，512，768，1024 の 12 段階で変化させた．評価指標は比較実験と同様， Accuracy と AUC を用い，10 回の試行を行った時の平均値を用いた．実験の結果を折れ線グラフにより図 3 に示す． NS では上限次元数が 16， Power では，128，Router では 32 で最も良い精度が得られた．データセットによって最高精度を達成するパラメータは異なっている．そのため，データセットごとにパラメータチューニングする必要があると考えられる．一方，全般的な性質として上限次元数が大きいと精度が低下する傾向がみられた．これは，次元数が大きくなることで無駄な情報量が多くなり，学習において提案特徴量に過剰に適合してしてしまうためだと考えられる．

5 関 連 研 究

5.1 リンク予測

リンク予測に関する初期の研究では，似たノードがリンクで接続するという仮定の下， common-neighbor index（CN）[4] や Adamic-Adder index（AA） [5] というヒューリスティックが提案されている． CN は，共通する隣接ノードの数によってノードペアのスコアリングを行う． AA は，接続数が多い隣接ノードは注目リンクのスコアリングにあまり寄与しないと仮定し，スコアリングを行う．これらはリンク予測対象のノードペアから 2 ホップ先までを活用する．そして， 2 ホップ先より長距離の情報を活用できる手法に Katz index [6] や PageRank [7] などがある． Katz index は，ノードペアに対する全ての経路のホップ数の合計でスコアリングする． PageRank は，片方のノードからもう片方のノードへランダムウォークにより到達する確率でスコアリングを行う．しかし，これらのヒューリスティックは，グラフのタイプを事前に仮定する必要がある．また，ノードやエッジに特徴量を持つグラフでは，それら特徴量を学習に活用することができないという欠点がある．これらの欠点を解消した手法として， VGAE [8] や SEAL [3] といった GNN に基づくリンク予測アルゴリズムが提案され，ヒューリスティックな手法を上回る精度を実現している．また，最近

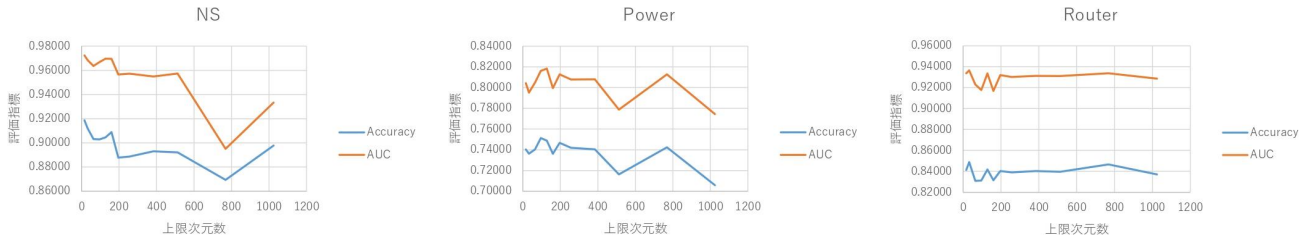


図 3: クラスタベクトルのパラメータである上限次元数によるリンク予測精度の変化．パラメータは 16，32，64，96，128，160，192，256，384，512，768，1024 の 12 段階で変化させた場合のリンク予測精度を示している．評価指標には Accuracy と AUC を用いる．データセットは左から順に NS， Power ， Router となっている．

SEAL の精度を上回る Walk Pool [10] という手法が提案されている． WalkPool では，ノード表現に attention を適用し，attention の共依存性をグラフのランダムウォークの遷移確率として解釈することを行う．これにより，ノード表現とグラフ構造の情報を，ある有効な潜在グラフ上のランダムウォーク遷移確率にエンコードし，それを用いて walk profiles という特徴量を計算することで高次の構造情報と抽出する．この手法は homophilic な（平均クラスタリング係数の値が高い）グラフと heterophilic な（平均クラスタリング係数の値が低い）グラフの両方のデータセットで安定した性能を発揮し，すべてのデータセットで最先端の性能を達成している．

5.2 グラフクラスタリング

まず，階層型グラフクラスタリングは，トップダウンアプローチとボトムアップアプローチの二種に大別することができる．トップダウンアプローチとは，グラフ全体を一つの大きなクラスタとして始め，リンクを切り離していくことで徐々に小さなクラスタへと分割していくアプローチである．一方，ボトムアップアプローチとは，各ノードを個々の独立したクラスタとして始め，一定の条件下でコミュニティを集約していくことでクラスタを得るアプローチである．

トップダウンアプローチとしての初期の手法である Normalized Cut 法 [17] は，クラスタ間のリンクが最小，クラスタ内のリンクが最大となるようにグラフを分割していき，クラスタを得る手法である．問題点として，時間計算量がノード数 n に対して，$O(n^3)$ となることや，事前にクラスタの個数を決定する必要があること，ノード数が均等なクラスタを形成できないことが挙げられる．

次に， Modularity ベースの手法について説明する．Girvan-Newman 法 [18] は， Modularity を評価することにより，グラフから削除するリンクを選択し，クラスタを順次形成していくトップダウンアプローチの手法である．この手法についても Normalized Cut 法と同様に計算量が $O(n^3)$ と非常に大きいことが問題点として挙げられている．1 時間当たりの処理データ量は 1 万ノード程度である．Newman [19] 法は，二つのノードを同一クラスタとした際に生じる Modularity の上昇量を計算し，その上昇量が最大となるノードの組合せを貪欲法を用いて探索を行い，クラスタを形成する手法であり，ボトムアップアプローチの手法である．また，この手法に対して，ヒープの導入やヒューリスティックを用いることで更なる高速化を実現した CNM [20] 法が提案されている．これらの手法は計算量が $O(n^2)$ へと改善されており，1 時間当たりの処理データ量は数百万ノード程度であり，大きく高速化が進んだことがわかる．Louvain 法 [1] は， 2 章で説明した手法であり，二つのフェーズを Modularity の向上が終了するまで反復する手法である．1 時間当たりの処理データ量は 1 千万ノード程度となり，さらに高速化が進んでいる．Incremental Agrregation 法 [21] はグラフの Power-Law 特性とグラフのクラスタ性を考慮することにより， Louvain 法と同程度の性能を保ちつつ，10 倍以上高速に計算できる手法である． Power-Law 特性とは大多数のノードは低次数で，少数のノードのみが高次数となるという実世界のグラフデータの性質のことである．この手法では，1 時間当たりの処理データ量は，一億から数十億ノードと，非常に高速な手法となっている．

5.3 ノードエンベディング

ノードエンベディングの学習手法には様々なものが提案されている． DeepWalk [22] は，グラフ上でのランダムウォークに現れるノードの分布が自然言語の単語の分布と似ていることから，単語表現モデルである Skip-Gram [23] に対して，ランダムウォークにより得られたノードの系列を適用することでノードエンベディングを得る手法である． Node2Vec [13] は，DeepWalk を改良したアルゴリズムであり，二つのパラメータの値を変えることにより，さまざまな粒度のノードエンベディングを獲得することができる． LINE [24] は，ノード同士がリンクでつながっているという一次の近接性と，同じノードを共有しているという二次の近接性に基づき，ノードエンベディングを学習する手法である． GraRep [25] では，1〜k 次の近隣情報を捉える表現を結合し， DeepWalk や LINE よりも大域的な情報を考慮したノードエンベディングを獲得することができる手法である． SDNE [26] は LINE と同様，一次の近接性と二次の近接性を考慮するが， Autoencoder の構造に基づいて学習を行うことで，近隣ノードの非線形変換によりノードエンベディングを獲得する手法である．また，両近接性を同時の考慮するため，一つのネットワークを用いて学習することが可能である．しかし，これらのノードエンベディング手法は，局所的な構造を捉えることができるが，コミュニティ構造はほとんど無視されている．それを受けて，コミュニティ構造をノー

ドエンベディングに埋め込む手法として M-NMF [27] という手法が提案されている．この手法では，NMF に基づくエンベディング学習と Modularity に基づいたコミュニティ検出モデルを同時に最適化することにより，局所的な構造と大域的な構造の両方を保持したノードエンベディングを獲得することができる．

6 終わりに

本稿では，大域的なグラフ構造を活用する部分グラフベースのリンク予測手法を提案した．提案手法では，大域的なグラフ構造の情報を得るために，階層型クラスタリングを行い，そのクラスタリング結果から，最小クラスタサイズとクラスタベクトルの二つの特徴量を作成し，既存のリンク予測手法である SEAL にそれらをそれぞれ付与して学習を行った．実験では，クラスタベクトルを用いた場合，五つのうちデータセットのうち四つにおいて，提案手法が既存手法を上回る精度を達成することを確認した．

今後の課題としては，クラスタベクトルについて，本稿では単純に用いる次元を選択することで次元削減を行っていたが，PCA [16] などの次元削減手法を用いることで，情報欠損を抑えることが挙げられる．また，これを用いることでパラメータチューニングの必要がなくなると考えられる．また，提案特徴量の代わりに，コミュニティ構造を明示的に捉えることができる M-NMF をノードエンベディングに用いた場合に，提案手法と精度を比較する必要があると考えられる．

謝　辞

本研究は JSPS 科研費 JP20H00583 の助成を受けたものです．

文　献

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, Vol. 2008, No. 10, 2008.

[2] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the ICLR*, 2017.

[3] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Proceedings of the NeurIPS*, 2018.

[4] M. E. J. Newman. Clustering and preferential attachment in growing networks. *The Physical Review E*, Vol. 64, No. 2, 2001.

[5] Adamic Lada, A and Adar Eytan. Friends and neighbors on the web. *The Social Networks*, Vol. 25, No. 3, 2003.

[6] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, Vol. 18, No. 1, 1953.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *The Computer Networks*, Vol. 30, No. 1, 1998.

[8] Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *arXiv*, 2016.

[9] Muhan Zhang and Yixin Chen. Weisfeiler-lehman neural machine for link prediction. In *Proceedings of the KDD*, 2017.

[10] Liming Pan, Cheng Shi, and Ivan Dokmanić. Neural link prediction with walk pooling. *arXiv*, 2021.

[11] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘ small-world ’ networks. *nature*, Vol. 393, No. 6684, 1998.

[12] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI*, 2018.

[13] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the SIGKDD*, 2016.

[14] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, Vol. 103, No. 23, 2006.

[15] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, Vol. 30, No. 7, 1997.

[16] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, Vol. 2, No. 11, 1901.

[17] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 22, No. 8, 2000.

[18] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, Vol. 69, No. 2, 2004.

[19] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, Vol. 69, No. 6, 2004.

[20] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, Vol. 70, No. 6, 2004.

[21] Hiroaki Shiokawa, Yasuhiro Fujiwara, and Makoto Onizuka. Fast algorithm for modularity-based graph clustering. In *Proceedings of the AAAI*, 2013.

[22] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the SIGKDD*, 2014.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013.

[24] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the WWW*, 2015.

[25] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the ACM*, 2015.

[26] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the SIGKDD*, 2016.

[27] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the AAAI*, 2017.