DEIM Forum 2019 F5-5

隣接性と構造類似性を考慮したグラフクラスタリング

小川　裕也† 　前川　政司†† 　竹内　孝††† 　佐々木勇和†† 　鬼塚　真††

† 大阪大学　〒 565–0871 大阪府吹田市山田丘 2–1
†† 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
††† NTT コミュニケーション科学基礎研究所　京都府相良郡精華町光台 2–4
E-mail: †{ogawa.yuya,maekawa.seiji,ogawa.yuya,onizuka}@ist.osaka-u.ac.jp, ††takeuchi.kou@lab.ntt.co.jp

あらまし　様々な分野でグラフデータが扱われており，グラフのコミュニティを発見することができるグラフクラスタリングは重要なタスクとなっている．本研究では隣接性と構造類似性を同時に考慮した新しいグラフクラスタリング手法である GCAS(Graph Clustering that leverages Adjacency and Structural similarity) を提案する．提案手法では隣接性または構造類似性だけを考慮することでは得られない，より正確なコミュニティを得ることが可能である．実際に実グラフデータでクラスタリング結果の正確性評価実験を行い，隣接性と構造類似性を同時に考慮することでクラスタリングの正確性が向上することを示した．

キーワード　グラフクラスタリング，構造類似度，行列分解

1 はじめに

グラフデータは人物や物事をノード，それらのつながりをエッジで表現したデータであり，社会学，生物学，コンピュータ科学などの様々な分野で幅広く用いられている [1]．代表的なグラフデータとしてソーシャルグラフが挙げられる．ソーシャルグラフは人と人とのつながりをグラフで表現したものであり，例えば Twitter[1] や Facebook[2] などの SNS 上でのつながりをグラフとして表現したものである．こういったグラフデータを解析できる技術のひとつにグラフクラスタリングがある．グラフクラスタリングは，グラフからコミュニティを抽出する目的で利用される．そして，グラフからコミュニティを発見することで，グラフの解析や様々なアプリケーションに利用が可能となる．そのため，グラフデータを扱う社会学，生物学やコンピュータ科学などの分野でグラフクラスタリングは重要な技術であり，需要が高まっている．実際に，生物学の分野ではタンパク質間相互作用ネットワークをクラスタリングすることでタンパク質の解析 [2] に用いている．また Sachan ら [3] によると，コミュニティを発見することでウェブマーケットやユーザーに対する効率的なウェブ広告の表示の役に立つと述べている．これらの例のように，グラフクラスタリングはグラフからコミュニティを抽出するために様々なことに活用されている．よって，より正確なコミュニティを抽出するために Modularity に基づく手法 [4–6] や min-max cut に基づく手法 [7,8] などの様々なグラフクラスタリング手法が提案されてきた．

グラフクラスタリングを行う際，重要な指標となるのは，ノード間の類似度である．そして，ノード間の類似度を考えるとき，重要な性質として隣接性と構造類似性がある．隣接性は，ノード同士が隣接しているかどうかを表す．実際のグラフデータにおいて，コミュニティ間のノードは隣接していることが多いが，コミュニティ外のノードとは隣接していることが少ない．つまり，隣接しているノード同士は同じコミュニティに属している可能性が高い．このことを利用して，[4–8] ではコミュニティ内ではエッジが多く，コミュニティ間ではエッジが少なくなるようなコミュニティを抽出する．一方，構造類似性は，ある 2 つのノードのそれぞれの隣接しているノード集合の類似性を表す．コミュニティ内のノード同士は共通して隣接しているノードが多く，構造類似性が強い．例えば，Structural Clustering Algorithm for Networks (SCAN) [9] や graph-Skeleton based Clustering (gSkeletonClu) [10] ではコミュニティ内で構造類似度が高くなるようなコミュニティを抽出する．このように既存手法では，隣接性と構造類似性が重要な性質であると認識されている．しかし，実際のグラフにおいてコミュニティ間でエッジが存在したり，コミュニティ内でエッジが十分に存在しないことが多く，隣接性もしくは構造類似性だけを考慮してクラスタリングを行うとうまくコミュニティを得ることができない問題がある．

そこで，本研究では隣接性と構造類似性を同時に考慮する新しいグラフクラスタリング手法 GCAS(Graph Clustering that leverages Adjacency and Structural similarity) を提案する．提案手法では隣接性を考慮するための隣接行列と構造類似性を考慮するための構造類似度行列の 2 つの行列を元に因子行列を得てクラスタリングを行う．これにより隣接性と構造類似度性を同時に考慮することを可能としている．また，構造類似性を考慮するにあたってノード間の構造類似度 [3] に Adamic/Adar 類似度を用いている．この類似度は SCAN における構造類似度とは異なり，次数が低いノードとのつながりを重視することでより的確な類似度の計算が可能となる．実際に，実グラフデー

1：http://www.twitter.com
2：http://www.facebook.com
3：本稿では構造類似度として狭義には SCAN における構造類似度を指す場合，広義には一般的に 2 つのノードの隣接ノード集合の類似度を指す場合がある

タを用いクラスタリングの正確性評価実験を行い，主に以下の結果を確認した．

- 隣接性と構造類似性を同時に考慮することでクラスタリングの正確性が向上すること
- 提案手法において Adamic/Adar 類似度が SCAN における構造類似度に比べ，より適切に構造類似度を表現できること

本稿の構成は以下のようになっている．まず 2 章で事前準備について述べる．ここでは本稿でのグラフの定義やクラスタリングの定義，そして提案手法で用いる既存研究について述べる．次に 3 章では提案手法について詳細に述べる．そして，4 章で評価実験及びその実験結果について述べる．5 章では関連研究について述べて，最後の 6 章にてまとめと今後の課題について述べる．

2 事前準備

グラフは $G=(V,E)$ として定義され，$V=\{v_0,v_1,...,v_{N-1}\}$ はノード集合，$E=\{(v_i,v_j)|0 \leqq i,j \leqq N-1, i \neq j\}$ はエッジ集合を表す．グラフクラスタリングはグラフ G が与えられたときグラフの構造情報からコミュニティ集合 $C=\{c_0,c_1,...,c_{K-1}\}$ を見つけ出す技術である．ただし，コミュニティ c_k はノード集合である．また，それぞれのノードはただ一つのコミュニティに属すものとする．つまり，$c_i \cap c_j = \emptyset (i \neq j)$ である．以下提案手法で用いている既存技術について述べる．

2.1 SymNMF

ここでは，提案手法のベースとなっている Kuang らによって提案されたグラフクラスタリング手法である SymNMF [11,12] について述べる．ノード $\mathcal{V}_i$ がコミュニティ $\mathcal{C}_k$ に対しどれだけ属しているかを表す行列 $\boldsymbol{V}=\{\boldsymbol{V}_{ik}\} \in \mathbb{R}_+^{N \times K}$ を考える．N はノードの数，K はコミュニティの数である．このとき行列 $\boldsymbol{A}=\boldsymbol{V}\boldsymbol{V}^T$ は行列 $\boldsymbol{X}$ から期待されるノード間類似度行列であると考えることができる．つまり，逆にノード間類似度行列 $\boldsymbol{A}$ が与えられたとき $\boldsymbol{A} \simeq \boldsymbol{V}\boldsymbol{V}^T$ のように行列 $\boldsymbol{V}$ で近似することでノード $\mathcal{V}_i$ がコミュニティ $\mathcal{C}_k$ に対しどれだけ属しているかを表す行列 $\boldsymbol{V}$ を得ることが可能である．SymNMF では以下の式 (1) を最適化することでノード間類似度行列 $\boldsymbol{A}$ から行列 $\boldsymbol{V}$ を得る．

$$\min_{\boldsymbol{V} \geqq 0} ||\boldsymbol{A}-\boldsymbol{V}\boldsymbol{V}^T||_F^2 \tag{1}$$

上式において $\boldsymbol{A} \in \mathbb{R}_+^{N \times N}$，$\boldsymbol{V} \in \mathbb{R}_+^{N \times K}$，$\mathbb{R}_+$ は非負の実数集合，$||\cdot||_F$ はフロベニウスノルムを表す．

最適化することで得た行列 $\boldsymbol{V}$ の各要素 $\boldsymbol{V}_{ik}$ は，ノード $\mathcal{V}_i$ がコミュニティ $\mathcal{C}_k$ に対しどれだけ属しているかを表すと考えることができる．よって，一番強く属しているコミュニティ $\mathcal{C}_j$ に属するものとする．ただし，$j=\arg\max_{1 \leqq k \leqq K} \boldsymbol{V}_{ik}$ である．SymNMF ではグラフの隣接行列 $\boldsymbol{S}$ が与えられたときこれをノード間類似度行列の一種であると考えノード間類似度行列 A の代わりに隣接行列 $\boldsymbol{S}$ を用いることで隣接性からコミュニティを得ることで可能となる．また，隣接行列ではなく構造類似度行列を入れることにより隣接性ではなく構造類似性からコミュニティを得ることも可能である．

2.2 構造類似度

2.2.1 SCAN における構造類似度

Xu らによって提案された密度ベースのグラフクラスタリング手法 SCAN [9] は，構造類似度に基づいてクラスタリングを行う．SCAN では，2 つのノード x と y の構造類似度 σ を以下の式 (2) のように定義している．

$$\sigma(x,y)=\frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)||\Gamma(y)|}} \tag{2}$$

ただし，$N(x)$ はノード x の隣接ノード集合であり，$\Gamma(x)=N(x) \cup \{x\}$ である．また，$|\Gamma(x)|$ は集合 $\Gamma(x)$ の要素数を表す．σ は 2 つのノードの共通隣接ノード集合の要素数を 2 つのノードそれぞれの隣接集合の要素数で正規化したものである．SCAN では，core と呼ばれるノードを中心に構造類似度 σ が高いノード同士は同じコミュニティであると考えることで，次々とコミュニティを拡大していき最終的なコミュニティ集合を得る．また，SCAN ではコミュニティに属さないノードをハブや外れ値として抽出する．

2.2.2 Adamic/Adar 類似度

2 つのノードを x と y，そしてそれらの隣接集合を $N(x)$，$N(y)$ としたとき，x と y の構造類似度 AAsim は以下のように表される．

$$\mathrm{AAsim}(x,y)=\sum_{u \in N(x) \cap N(y)} \frac{1}{log(|N(u)|)} \tag{3}$$

ノード u は，ノード x とノード y の共通隣接集合 $N(x) \cap N(y)$ に含まれるノードであり，$|N(u)|$ はノード u の隣接ノード集合の数を表す．この Adamic/Adar 類似度は，2 つのノードの共通隣接ノード集合が多いほど高くなるという点において，上で述べた SCAN で定義された構造類似度と同じである．しかし，Adamic/Adar 類似度では共通隣接ノードの内，次数が高いノードより次数が低いノードの方により大きな重みが与えられる．論文引用関係を表すグラフを例に挙げると，他の論文からも引用が多い論文を共通して引用していても似ているか判断しにくいが，比較的他の論文からも引用が少ない論文を共通して引用していると似ていると判断しやすいことを表す．この考え方は自然な考え方であり, すべてのノードを同じ重みで扱うよりも現実に即した類似度であると考えられる．

3 提案手法

本稿では，隣接性と構造類似性を同時に考慮した新しいグラフクラスタリング手法である GCAS(Graph Clustering that leverages Adjacency and Structural similarity) を提案する．隣接性と構造類似性という 2 つの観点を組み合わせてコミュニティを得ることで，どちらか一方だけでは得ることができないより正確なコミュニティを抽出することが可能になる．例えば，

隣接はしているが構造類似度は高くないノードや，構造類似度は高いが隣接はしていないノードについて，隣接性と構造類似性を同時に考慮することで互いに情報を補完してコミュニティを得ることができる．また，提案手法では構造類似性を考慮するにあたって，2 章で述べた理由によりノード間の構造類似度として SCAN における構造類似度ではなく Adamic/Adar 類似度を用いることで，より正確なコミュニティを抽出することが可能である．

提案手法では，隣接性と構造類似性を同時に考慮するために隣接行列 S と構造類似度行列 W の 2 つの行列を用いて以下の式 (4) で表す目的関数を最適化することで，因子行列 V を得る．式 (4) において第一項で SymNMF と同様に隣接性から因子行列 V を得て，第二項で構造類似性によって因子行列 V の行ベクトル v_i に制約を加えて，構造類似度が高いノード同士に対応する行ベクトルの距離を近づけている．素子て，最適化することで得た行列 $\boldsymbol{V}$ の各要素 $\boldsymbol{V}_{ik}$ は，ノード $\mathcal{V}_i$ がコミュニティ$\mathcal{C}_k$ に対しどれだけ属しているかを表すと考えることができるため，一番強く属しているコミュニティ$\mathcal{C}_j$ に属するものとする．ただし，$j = \arg\max_{1 \leqq k \leqq K} \boldsymbol{V}_{ik}$ である．以下に，提案手法における目的関数を示す．

$$\min_{V \geqq 0} ||S - VV^T||_F^2 + \lambda \sum_{ij} W_{ij} ||v_i - v_j||^2 \tag{4}$$

ただし，$S \in \mathbb{R}_+^{N \times N}$，$V \in \mathbb{R}_+^{N \times K}$，$W \in \mathbb{R}_+^{N \times N}$，$\mathbb{R}_+$ は非負の実数集合，$||\cdot||_F$ はフロベニウスノルム，$\lambda \geqq 0$ はハイパーパラメータであり，隣接性と構造類似性による制約を統合する重みを表す．また，N はノードの数,K はコミュニティの数を表す．隣接行列 S は，その要素 s_{ij} がノード v_i と v_j にエッジがあれば 1 なければ 0 となる行列であり，構造類似度行列 W は，その要素 w_{ij} が $w_{ij} = \mathrm{AAsim}(v_i, v_j)$ となる行列である．

3.1 更 新 規 則

ここでは，提案手法の更新規則について述べる．まず，式 (4) の目的関数を損失関数として書き換えると以下の式 (5) となる．

$$L = ||S - VV^T||_F^2 + 2\lambda Tr(V^T L V) + Tr(\alpha^T V) \tag{5}$$

ここで，α は $V \geqq 0$ の制約のためのラグランジュ未定乗数行列である．また $\sum_{ij} W_{ij} ||v_i - v_j||^2 = 2Tr(V^T L V)$ であることを用いた．ただし $L = D - W$ であり，行列 D は $D_{ii} = \sum_j W_{ij}$ である対角行列である．損失関数 L を V について偏微分すると以下のようになる．

$$\frac{\partial L}{\partial V} = -4SV + 4VV^T V + 4\lambda LV + \alpha \tag{6}$$

$L = D - W$ であるから

$$\frac{\partial L}{\partial V} = -4SV + 4VV^T V + 4\lambda DV - 4\lambda WV + \alpha \tag{7}$$

ここで，KKT 条件を考慮すると $\alpha_{ij} V_{ij} = 0, \frac{\partial L}{\partial V} = 0$ である．よって，以下の式 (8) の更新式を得る．

$$V_{ij}^{new} = V_{ij} \frac{(SV + \lambda WV)_{ij}}{(VV^T V + \lambda DV)_{ij}} \tag{8}$$

3.2 GCAS アルゴリズム

提案手法のアルゴリズムを Algorithm 1 で示す．GCAS では入力を隣接行列 S と構造類似度行列 W として，コミュニティ集合 C を出力する．初めに，行列 V を初期化し (詳しい手順は後で述べる)，反復規則に従って行列 V を反復更新する．更新を終えると，SymNMF と同様に行列 V からコミュニティ集合 C を得る．また，行列 V の初期化の手順を以下に示す．なお，kmeans 法 [13] ではなく kmeans++法 [14] を用いたのは一般に kmeans++法の方がクラスタリング性能が高いからである．

（1） 隣接行列を入力として k-means++法 でコミュニティ集合 C=$\{c_0, c_1, ..., c_K\}$ を得る

（2） 行列 V の要素のすべてを 0 とする

（3） ノード v_i がコミュニティc_k に属しているなら行列 V の要素 v_{ik} を 1 とする

（4） 3 の手順をすべてのノードについて行う

例えば，隣接行列を入力に k-means ++法 でクラスタリングすることで図 1 のようなコミュニティ集合 C が得られたなら，まず行列 V の要素すべてを 0 とした後，ノード v_0 がコミュニティc_0 に属しているため行列 V の要素 v_{00} を 1 とする．そしてノード v_1 がコミュニティc_0 に属しているため行列 V の要素 v_{10} を 1 とする．以下同様の操作を繰り返すと行列 V は以下のように初期化される．[4]

$$V = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{9}$$

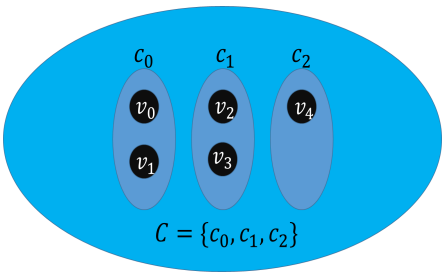

図 1　コミュニティ集合 C

3.3 計　算　量

この章では，提案手法の計算量について述べる．提案手法において行列 V の更新において $N \times N$ の行列と $N \times K$ の行列の積演算が行われる．この計算量は $O(N^2 K)$ となる．そしてこの行列の積演算が提案手法の計算量において支配的な項となるため，提案手法の計算量は $O(N^2 Kt)$ となる．ここで N はノードの数，K はコミュニティの数，t は学習の反復回数である．

4 : 実際には，行列の要素が 0 になると更新できなくなることがあるため，行列の各要素に $\epsilon \ll 1$ であるような非常に小さい数 ϵ を足している．

Algorithm 1 GCAS アルゴリズム

Input: $\boldsymbol{S}, \boldsymbol{W}$
Output: clustering result $\mathcal{C}$
Initialize $\boldsymbol{V}$
  set($\boldsymbol{V}$) $\leftarrow \boldsymbol{0}$
  $\mathcal{C} \leftarrow$ k-means++($\boldsymbol{S}$)
  set($\boldsymbol{V}$) $\leftarrow \mathcal{C}$
Learn $\boldsymbol{V}$
  for counter=1 to max iteration do
    $\boldsymbol{V}^{(t+1)} \leftarrow$ update($\boldsymbol{V}^{(t)}$)
  end do
Assign nodes to community
  $\mathcal{C} \leftarrow$ max_ entry($\boldsymbol{V}$)

3.4 GCAS における構造類似度

提案手法では構造類似度に Adamic/Adar 類似度を用いているが，Adamic/Adar 類似度の代わりに SCAN で用いられている構造類似度などの異なる類似度を用いることで Adamic/Adar 類似度を用いる場合と違ったクラスタリング結果が得られる．本稿では Adamic/Adar 類似度の代わりに SCAN で用いられている構造類似度を用いた GCAS を GCAS* と呼ぶこととする．

4 評価実験

この章では提案手法のクラスタリング正確性の評価を行う．また，パラメータのクラスタリング正確性への影響についても述べる．また，比較手法として以下の手法を用いた．[5]

- GCAS*：Adamic/Adar 類似度が構造類似度に適しているか確認するために，提案手法において Adamic/Adar 類似度ではなく SCAN における構造類似度を採用した手法．
- SymNMF：隣接性と構造類似性を同時に考慮することでクラスタリング結果の正確性が向上するか確認するために，隣接性もしくは構造類似性のうち一つだけを考慮する手法．入力によって以下のような表記を行う．
  - SymNMF_A：入力を隣接行列とした SymNMF．隣接性のみを考慮する．
  - SymNMF_S：入力を Adamic/Adar 類似度を用いた構造類似度行列とした SymNMF．構造類似性のみを考慮する．
  - SymNMF_S*：入力を SCAN における構造類似度を用いた構造類似度行列とした SymNMF．構造類似性のみを考慮する．
- k-means++法 [6]：広く用いられているクラスタリング手法．提案手法及び SymNMF における因子行列の初期化に利用している．本実験では入力を隣接行列としているため，隣接性のみ考慮する．
- node2vec [15]：ネットワークエンベディング手法．DeepWalk [16] の拡張手法であり，パラメータによって幅優先と深さ優先による探索の影響度を制御することが可能である．広義に，隣接性と構造類似性を考慮している．

4.1 データセット

データセットとして以下の表 1 に示すようなグラフデータを用いる．

表 1　データセット

データセット名	ノード数	エッジ数	クラスタ数
Parliament	451	5823	7
Cora	2708	5278	7
BlogCatalog	5196	171743	6
LFR1000	1000	8261	8

- Parliament [17]：フランス国会の議員の関係を表したグラフである．ノードは議員を表し，エッジは議案に連署したことを表す．正解コミュニティは各政党となっている．
- Cora[7]：機械学習に関する論文の引用関係を表したグラフである．ノードは論文を表し，エッジは引用関係を表す．正解コミュニティは，論文が属する以下の 7 つの機械学習のサブカテゴリになっている．Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, Theory.
- BlogCatalog [18]：ソーシャルブログである BlogCatalog[8] 上での，ユーザーの関係を表したグラフである．ノードはユーザーで，エッジは BlogCatalog 上で友達関係であることを表す．正解コミュニティは，BlogCatalog 上で登録する際にあらかじめ決めるカテゴリとなっている．
- LFR1000：LFRbenchmark [19] によって生成された人口グラフである．パラメータは [19] で使用されている値を参考に用いた．値は以下の通りである．n=1000, tau1=3.0, tau2=2.0, mu=0.5, average_degree=15,min_community=50. その他のパラメータはあらかじめ設定されている値を用いた．グラフと同時に出力されるコミュニティを正解コミュニティとした．

4.2 評価指標

クラスタリング結果の正確性の評価指標として，クラスタリングの評価指標として広く用いられている Adjusted Rand Index (ARI) 及び Normalized Mutual Information (NMI) を用いる．以下にこの 2 つの評価指標の詳細を述べる．

- ARI [20]：正解コミュニティと得られたコミュニティの一致度を測る指標であり，0 から 1 の値をとる．ARI が 1 に近いほど正解コミュニティと得られたコミュニティの一致度が高いことを示す．つまり ARI が高いほどクラスタリング結果の正確性が高いことを表す．ARI は以下のように定義される．

集合 $X = \{X_1, X_2, ..., X_r\}$ と集合 $Y = \{Y_1, Y_2, ..., Y_s\}$ を考え，集合 X_i と Y_j の共通集合 $X_i \cap Y_j$ の要素数を n_{ij} とする．そして $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$ であるとすると，ARI は以

5：SCAN ではコミュニティの数を決めることができず，コミュニティに属さないノードも出力されるため本稿では比較を行わない．
6：https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
7：https://linqs.soe.ucsc.edu/data
8：http://www.blogcatalog.com

下のように計算される．

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \mathrm{EI}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - \mathrm{EI}},\ \mathrm{EI} = \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}} \tag{10}$$

ただし $\binom{n}{k} = {}_n\mathrm{C}_k$ である．

- NMI : ARI と同様正解コミュニティと得られたコミュニティの一致度を測る指標であり．0 から 1 の値をとる．1 に近いほど正解コミュニティと得られたコミュニティの一致度が高いことを示す．正解コミュニティ集合を C，クラスタリングによって得られたコミュニティ集合 C' とすると NMI は [21] を参考にすると以下のように定義される．

$$\mathrm{NMI}(C, C') = \frac{\sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}}{\sqrt{H(C)H(C')}} \tag{11}$$

上の式において c_i, c'_i はそれぞれ正解コミュニティの内のひとつ，クラスタリングによって得られたコミュニティ内のひとつを表す．$p(c_i)$ と $p(c'_j)$ は，ランダムに選ばれたノードがそれぞれ c_i，c'_j に属する確率を表す．また $H(C)$，$H(C')$ はそれぞれコミュニティ集合 C，C' のエントロピーを表す．NMI も ARI と同様に高いほどクラスタリング結果の正確性が高いことを示す．

4.3 実験設定

Adamic/Adar 類似度の計算において，[22] では自然対数を用いているが，本稿ではより次数が低いノードを重視するため常用対数を用いて構造類似度の計算を行った．SymNMF の行列 V の初期化は，提案手法と同様に行った．node2vec は，[15] を参考にして，得られたノードの表現ベクトルを入力として k-means++によってクラスタリングを行った．node2vec のハイパーパラメータは p=1, q=0.5 として他の値はあらかじめ決められている値を用いた [9]．そして提案手法のハイパーパラメータ λ は { 0.01, 0.02,..., 0.09, 0.1 , ..., 0.9} の中から各データに対し最適なもので実験を行った．また ARI 及び NMI の値はすべての手法において実験を 10 回行いその平均値をその手法で得られた結果とした．括弧内の数字は標準偏差を表す．表 2，3 に結果を示す．

表 2　ARI の平均値．括弧内の数字は標準偏差を表す．*は構造類似度に σ を利用したことを表す．また太字はそれぞれのグラフデータに対して最も値が高いことを表す．

手法	Parliament	Cora	BlogCatalog	LFR1000
GCAS	**0.902**(0.081)	**0.339**(0.027)	**0.140**(0.012)	**0.199**(0.014)
GCAS*	0.866(0.052)	0.322(0.022)	0.136(0.013)	0.197(0.008)
SymNMF_A	0.584(0.058)	0.263(0.016)	0.132(0.023)	0.150(0.015)
SymNMF_S	0.582(0.010)	0.206(0.023)	0.000(0.000)	0.139(0.016)
SymNMF_S*	0.496(0.032)	0.245(0.040)	0.123(0.030)	0.129(0.025)
k-means++	0.279(0.049)	0.014(0.015)	0.021(0.004)	0.055(0.026)
node2vec	0.000(0.000)	0.223(0.004)	0.000(0.000)	0.000(0.000)

9 : https://github.com/aditya-grover/node2vec

表 3　NMI の平均値．括弧内の数字は標準偏差を表す．*は構造類似度に σ を利用したことを表す．また太字はそれぞれのグラフデータに対して最も値が高いことを表す．

手法	Parliament	Cora	BlogCatalog	LFR1000
GCAS	**0.875**(0.039)	**0.400**(0.018)	**0.231**(0.010)	**0.336**(0.012)
GCAS*	0.845(0.028)	**0.400**(0.013)	**0.231**(0.022)	0.287(0.013)
SymNMF_A	0.700(0.027)	0.345(0.018)	0.220(0.020)	0.256(0.017)
SymNMF_S	0.679(0.006)	0.308(0.016)	0.012(0.001)	0.193(0.024)
SymNMF_S*	0.673(0.023)	0.354(0.026)	0.212(0.026)	0.200(0.025)
k-means++	0.512(0.027)	0.136(0.057)	0.127(0.016)	0.100(0.029)
node2vec	0.027(0.000)	0.261(0.007)	0.001(0.000)	0.016(0.003)

4.3.1 クラスタリング結果の正確性評価

表 2，3 においてすべてのグラフデータにおいて隣接性と構造類似性を同時に考慮する提案手法 GCAS が最も良いクラスタリング結果を示していることが確認できる．このことから，隣接性と構造類似性を同時に考慮することでクラスタリング結果が向上することがわかる．また GCAS が GCAS* と同程度以上のクラスタリング性能を示していることから，すべてのノードを同じ重みで扱う SCAN の構造類似度よりも Adamic/Adar 類似度の方が構造類似度として適していることがわかる．よって，本稿で用いた 3 つのグラフデータについて，次数が高いノードよりも次数が低いノードと共通して隣接しているほうがより類似度が高いという Adamic/Adar 類似度における仮定が正しいことがわかる．

4.4 因子行列の初期化のクラスタリング結果への影響

提案手法において得られるコミュニティは因子行列 $\boldsymbol{V}$ の初期値に依存する．そこで，提案手法において因子行列 $\boldsymbol{V}$ を k-means++によって得られるコミュニティによって初期化しているが，ランダムな値で初期化した場合クラスタリング結果の正確性がどう変化するのか確かめる．そのため k-means++法によって因子行列を初期化した場合とランダムな値で初期化した場合の提案手法のクラスタリング結果の正確性を比較する．ランダムな値で初期化した提案手法を $\mathrm{GCAS}_{\mathrm{ran}}$ と表すこととする．以下表 4, 5 に結果を示す．

表 4　ARI の平均値．括弧内の数字は標準偏差を表す．また太字はそれぞれのグラフデータに対して最も値が高いことを表す．

手法	Parliament	Cora	BlogCatalog	LFR1000
GCAS	**0.902**(0.081)	**0.339**(0.027)	**0.140**(0.012)	**0.199**(0.014)
$\mathrm{GCAS}_{\mathrm{ran}}$	0.886(0.054)	0.313(0.031)	0.136(0.007)	0.194(0.025)

表 5　NMI の平均値．括弧内の数字は標準偏差を表す．また太字はそれぞれのグラフデータに対して最も値が高いことを表す．

手法	Parliament	Cora	BlogCatalog	LFR1000
GCAS	**0.875**(0.039)	**0.400**(0.018)	**0.231**(0.010)	**0.336**(0.012)
$\mathrm{GCAS}_{\mathrm{ran}}$	0.858(0.028)	0.387(0.016)	0.216(0.008)	0.328(0.020)

表 4, 5 からランダムな値で初期化するよりも k-means++によって初期化を行う方がクラスタリング結果が良いことがわか

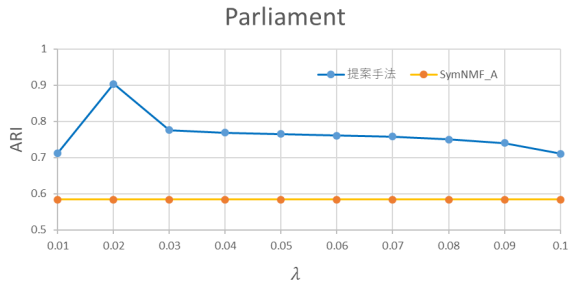

図 2　Parliament_ARI

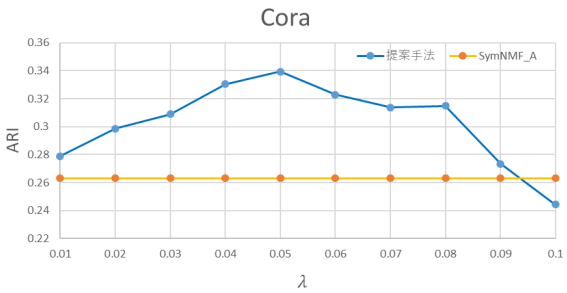

図 3　Cora_ARI

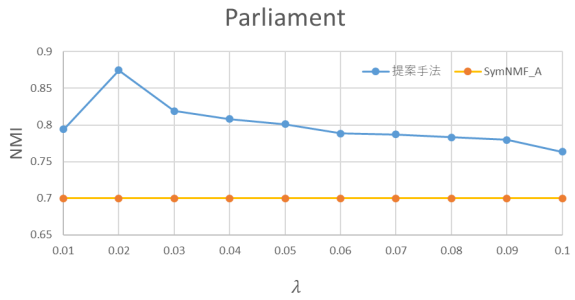

図 4　Parliament_NMI

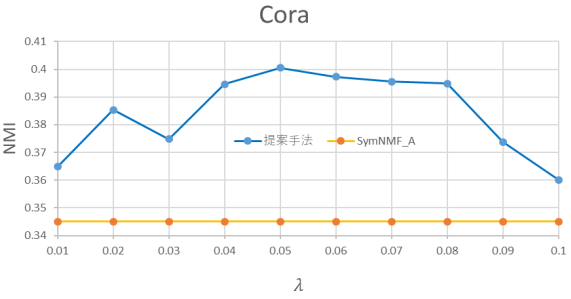

図 5　Cora_NMI

る．つまり，全くランダムな値から学習を進めるよりも，事前学習をしてから学習した方が良いクラスタリング結果が得られる．

4.5　バランスパラメータのクラスタリング正確性への影響

図 2 と図 3 に各グラフデータにおいてのバランスパラメータ λ と ARI の関係を示す．また図 4 と図 5 に各グラフデータにおいてのバランスパラメータ λ と NMI の関係を示す．図 3 の結果からバランスパラメータ λ の値を大きくすると提案手法が SymNMF_A よりもクラスタリング性能が悪くなることがあることがわかる．これはバランスパラメータ λ を大きすぎると，制約項による制約が強くなりうまくコミュニティが得ることができなくなるためであると考えられる．よって，本稿ではバランスパラメータ λ の値を大きくしすぎるとクラスタリング結果の正確性が下がる可能性を考慮して λ の値は 0.05 以下で使用することを推奨する．

5　関 連 研 究

5.1　GNMF

Graph regularized NMF(GNMF) [23,24] は NMF を改良してグラフ制約を課した行列分解手法であり，その目的関数は提案手法と関連が深い．各ノードに対する各特徴量を表現したデータ行列 X から，ノードのクラスタリングを行うために利用される．以下式 (12) に目的関数を示す．

$$\min_{V \geqq 0} ||X - UV^T||_F^2 + \lambda \sum_{ij} W_{ij} ||v_i - v_j||^2 \qquad (12)$$

式 (12) において $X \in \mathbb{R}_+^{M \times N}$，$U \in \mathbb{R}_+^{M \times K}$，$V \in \mathbb{R}_+^{N \times K}$，$W \in \mathbb{R}_+^{N \times N}$，$\lambda$ はハイパーパラメータである．また N はノードの数，M は属性の数，K はコミュニティの数を表す．行列 W は以下のように定義されている．

$$W_{i,j} = \begin{cases} 1 & \text{if} \quad \mathcal{V}_i \in N_p(\mathcal{V}_j) \quad \text{or} \quad \mathcal{V}_j \in N_p(\mathcal{V}_i) \\ 0 & \text{otherwise} \end{cases} \qquad (13)$$

$N_p(\mathcal{V}_i)$ はノード $\mathcal{V}_i$ の p 近傍ノード集合である．目的関数第 1 項においてデータ行列から特徴を抽出する．そして，目的関数第 2 項においてノード間の距離が近いと制約をかけることで，因子行列の行ベクトルの距離を近づける．

5.2　node2vec

node2vec [15] は，ネットワークエンベディング手法である．グラフデータから，ノードの表現ベクトルを得ることを目的に使用される．単語をベクトルに変換する技術である Word2vec をグラフに応用するために，ランダムウォークを用いてサンプリングを行う．DeepWalk [16] の拡張手法であり，パラメータによって幅優先と深さ優先による探索の影響度を制御することが可能である．node2vec では，得られたベクトルを元にして，主にリンク予測やマルチラベリングが行われる．DeepWalk や node2vec などのネットワークエンベディング手法におけるノードの表現ベクトルは，提案手法において因子行列の行ベクトルに対応する．つまり，ノードの表現ベクトルを得るという点では，関連が深い技術である．

6　お わ り に

本稿では隣接性に加え，Adamic/Adar 類似度を用いた構造類似性を考慮した新しいグラフクラスタリング手法を提案した．提案手法では隣接性と構造類似性を同時に考慮することで一方だけでは得ることができないより正確なコミュニティを得るこ

とを可能とした．

今後の課題としては，ノードの属性を利用できるように拡張，ノードの表現ベクトルが得られることを利用してクラシフィケーションタスクでの応用などが挙げられる．

文　献

[1] Santo Fortunato. Community detection in graphs. *Physics reports*, Vol. 486, No. 3-5, pp. 75–174, 2010.

[2] Sylvain Brohee and Jacques Van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, Vol. 7, No. 1, p. 488, 2006.

[3] Mrinmaya Sachan, Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam. Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pp. 331–340. ACM, 2012.

[4] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, Vol. 69, No. 6, p. 066133, 2004.

[5] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, Vol. 70, No. 6, p. 066111, 2004.

[6] 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, p. P10008, 2008.

[7] Chris HQ Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst D Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 107–114. IEEE, 2001.

[8] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 22, No. 8, pp. 888–905, 2000.

[9] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas AJ Schweiger. Scan: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 824–833. ACM, 2007.

[10] Heli Sun, Jianbin Huang, Jiawei Han, Hongbo Deng, Peixiang Zhao, and Boqin Feng. gskeletonclu: Density-based network clustering via structure-connected tree division or agglomeration. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 481–490. IEEE, 2010.

[11] Da Kuang, Chris Ding, and Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pp. 106–117. SIAM, 2012.

[12] Da Kuang, Sangwoon Yun, and Haesun Park. Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering. *Journal of Global Optimization*, Vol. 62, No. 3, pp. 545–574, 2015.

[13] Pankaj K Agarwal and Nabil H Mustafa. K-means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 155–165. ACM, 2004.

[14] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[15] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

[16] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710. ACM, 2014.

[17] Aleksandar Bojchevski and Stephan Günnemann. Bayesian robust attributed graph clustering: Joint learning of partial anomalies and group structure. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] Xiao Huang, Jundong Li, and Xia Hu. Accelerated attributed network embedding. In *SIAM International Conference on Data Mining*, pp. 633–641, 2017.

[19] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, Vol. 78, No. 4, p. 046110, 2008.

[20] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, Vol. 2, No. 1, pp. 193–218, 1985.

[21] Shudong Huang, Zenglin Xu, and Fei Wang. Nonnegative matrix factorization with adaptive neighbors. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pp. 486–493. IEEE, 2017.

[22] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, Vol. 25, No. 3, pp. 211–230, 2003.

[23] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Nonnegative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pp. 63–72. IEEE, 2008.

[24] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 8, pp. 1548–1560, 2011.