情報処理学会論文誌 Vol.58 No.1 278–287 (Jan. 2017)

滑らかな変化を検出するためのグラフ系列クラスタリング

長村 佳歩[1,†1] 奥井 颯平[2] 猪口 明博[1,a)]

受付日 2016年4月19日, 採録日 2016年10月4日

概要：本稿では，時間とともに変化するグラフを対象として，互いに結び付きの強いクラスタ，およびその変化を検出する問題を扱う．Preserving Cluster Membership と呼ばれる従来手法では，検出されるクラスタリングの結果は，はじめに与えられるグラフのクラスタリングのしやすさに依存し，いかなる時刻のグラフもちょうど k 個にしか分割できないという課題をかかえていた．本稿では，これらの課題を克服し，スペクトラルクラスタリングに基づいた手法を提案する．提案手法は，各時刻のグラフを k，あるいはそれ以下のクラスタに分割することができるため，時間とともに分裂するクラスタ，あるいは併合していくクラスタを検出することが可能である．

キーワード：クラスタリング，グラフ系列，スペクトラルクラスタリング，グラフカット

Graph Sequence Clustering for Detecting Smooth Changes

Kaho Osamura[1,†1] Sohei Okui[2] Akihiro Inokuchi[1,a)]

Received: April 19, 2016, Accepted: October 4, 2016

Abstract: This paper tackles the problem of detecting clusters in which vertices strongly connect with one another in an evolving graph. The conventional method, called Preserving Cluster Membership, is based on spectral clustering and has some drawbacks: the result detected by the method depends on the ease of clustering the initial graph and the graph is partitioned into exactly k clusters. In this paper, we propose a graph clustering method to overcome the drawbacks of using a spectral clustering approach. The proposed method detects clusters, some of which divide or merge over time because the method can partition a graph into k or fewer clusters.

Keywords: clustering, graph sequence, spectral clustering, graph cut

1. はじめに

情報技術の発展により，膨大な量のデータを蓄積することが可能となった．しかし，日々肥大化するデータは人間の理解力を超えたため，有益な情報が含まれていてもそのままでは理解できなくなっている．そこで，そのような膨大なデータから有益な情報を発見するため，近年データマイニングに関する研究が注目され，さかんに研究されている．そのなかでもクラスタリングは，教師なし学習手法であるため，カテゴリが未知のデータを分類するのに有用である．

本研究が対象とするグラフ系列マイニングでは，グラフ系列を複数のクラスタ系列に分割する．ここでグラフ系列とは，辺の重みが時間変化するグラフを時刻順に並べたものである．またクラスタとは，各時刻のグラフの頂点を辺の結び付きが強いものどうしがまとまるように分割したものである．さらに，クラスタ系列とは，各時刻のグラフの頂点集合を分割してできた複数のクラスタから 1 つを選び，時刻順に並べたものであり，クラスタの時間変化を表したものである．たとえば，人間関係ネットワークにおいて，人をグラフの頂点，人と人の関係をグラフの辺で表し，その親密度に応じて辺を重み付けすると，ある時点の人間

[1] 関西学院大学理工学部
School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo 669–1337, Japan
[2] 関西学院大学大学院理工学研究科
Graduate School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo 669–1337, Japan
[†1] 現在，奈良先端科学技術大学院大学
Presently with Nara Institute of Science and Technology
[a)] inokuchi@kwansei.ac.jp

関係ネットワークを重み付きグラフにより表現することができる．さらに時刻の経過とともにその構造が変化する人間関係ネットワークは，重み付きグラフの系列として表すことが可能である．また，人間関係ネットワークでは親密度が大きい人同士が集まってクラスタが形成され，クラスタの時間変化を表すクラスタ系列は，クラスタに含まれる人の人間関係の変化を表す．人間関係ネットワークを表すグラフ系列をクラスタ系列に分割することにより，人間関係ネットワークに隠れたクラスタの変化を発見することが期待され，マーケティング戦略などへ役立てることができると考えられる．

2. 関連研究

本稿では，グラフの辺の重みが時間変化するデータを対象とする．文献 [16] において，特徴空間上の事例集合の 2 事例間の距離の逆数を辺の重みとすることで，特徴空間上の事例集合をグラフに変換できることが紹介されている．そこで以下では，文献に従い [17]，特徴空間上の事例集合のクラスタリングを扱ったクラスタリング問題を関連研究として紹介する．

時刻 t における事例集合を X_t とし，T ステップからなる X_t の系列を $D = \langle X_1, X_2, \cdots, X_T \rangle$ で表す．D をクラスタリングする問題を 4 つのタイプに分類することができる．第 1 タイプは，**図 1** に示すようにデータストリームとして時々刻々と届く事例をクラスタリングする問題 [1], [2], [3], [6], [8], [10], [14] である．ただし，1 時刻ステップに届くデータは 1 事例であるものとする．このタイプの手法はオンラインデータ解析の問題である．

第 2 のタイプは，図 1 に示すように n 本の系列データを k 個のクラスタに分割する問題 [4], [13], [15] であり，このタイプの問題はバイオインフォマティクス分野の DNA 配列のクラスタリング問題やタンパク質配列のクラスタリング問題と同様である．

第 3 のタイプは n 本のデータストリームを k 個のクラスタに分割する問題 [5], [7] である．オンラインデータ解析である点は，第 1 のタイプと同じであるが，ある時刻で n 個の事例集合を扱う点が異なる．このタイプの問題は，PCM

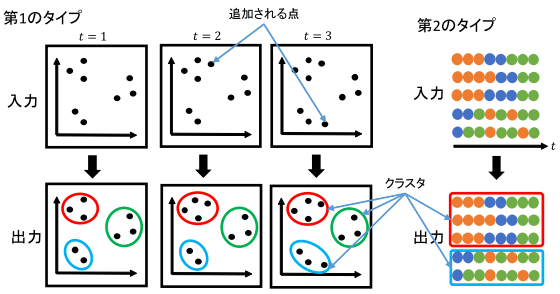

図 1 $\langle X_1, X_2, \cdots, X_T \rangle$ をクラスタリングする問題のタイプ
Fig. 1 Problem types for clustering $\langle X_1, X_2, \cdots, X_T \rangle$.

(Preserving Cluster Membership) と呼ばれる手法で扱われている問題であり，3.2 節で図を用いながら詳細に紹介する．

第 4 のタイプは本研究が扱う問題 [12], [17] である．第 1 のタイプと異なり，n 個の事例集合からなるデータ X_t の系列であり，無限のデータストリームはなく T ステップからなるデータ D は事前に与えられている．第 2 のタイプではクラスリングされる個々のデータは系列データであるが，このタイプではクラスリングされる個々のデータは X_t の各事例である．また，第 2 のタイプでは k 個のクラスタからなる 1 つの結果が出力されるのにされるのに対して，このタイプでは k 個のクラスタからなる結果が T 個出力される．第 3 のタイプではある時刻 t においてそれ以前のデータを用いて X_t をクラスタリングするのに対して，このタイプでは X_t をクラスタリングするためにその前後の時刻のデータを用いる．

クラスタリング問題は，m 次元空間の事例集合 $D' = \{\vec{x}_i \mid \vec{x}_i \in \mathcal{R}^m, i \in [1,n]\}$ を k 個のクラスタに分割する問題である．k-means では，互いに素な k 個のクラスタ $\{C_1, C_2, \cdots, C_k\}$ に分割される．ただし，$D' = \bigcup_{j=1}^{k} C_j$ である．k-means は，

$$\sum_{j=1}^{k} \sum_{\vec{x}_i \in C_j} ||\vec{x}_i - \vec{c}_j||^2$$

を最少化する問題として定式化される．ここで，$\vec{c}_j$ は C_j の重心である．k-means では，各事例は 1 つのクラスタのみに属するという制約があるが，Fuzzy c-means では，$\vec{x}_i$ の C_j への帰属率 $\mu_{ij} \in [0,1]$ を与え，

$$\sum_{j=1}^{k} \sum_{i=1}^{n} (\mu_{ij})^p ||\vec{x}_i - \vec{c}_j||^2$$

を最少化する問題として定式化される．ここで，$p > 1$ はべき乗パラメータである．また，k-means は，外れ値があるデータに対して頑健でないという欠点がある．これを克服する手法としてスペクトラルクラスタリングがあり，3 章で扱う．データの生成モデルとして，混合ガウス分布 $\sum_{j=1}^{k} \pi_j N(\vec{\mu}_j, \Sigma_j)$ を与え，そのパラメータ $\vec{\mu}_j$ と Σ_j を D から推定する問題として定式化する問題もある．このほか，階層型クラスタリングと呼ばれる手法もある．しかし，これらの手法は時間とともに変化するグラフに対して，直接適用できる手法ではない．

3. グラフ系列のクラスタリング

3.1 問題定義

時刻 t のグラフを $G^{(t)} = \left(V, E, w^{(t)}\right)$ で表す．ただし，V は頂点の集合であり，その要素には個別の ID が振られており一意に識別可能であるとする．また，E はすべての辺の集合 $E = V \times V$ であり，$w^{(t)}$ は時刻 t において辺に非負の実数値を割り当てる関数である．このような T 個のグ

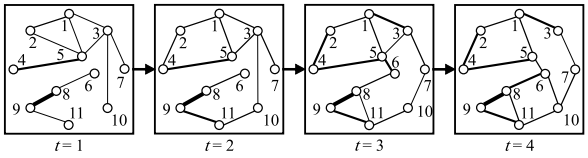

図 2 4 ステップ重み付きグラフ系列の例（頂点に付随する数字は頂点 ID を表す）

Fig. 2 Example of a weighted graph sequence with four steps (numbers attached to vertices represent vertex IDs).

ラフを時系列順に並べたものを，T ステップ重み付きグラフ系列とし，$\langle G^{(1)}, G^{(2)}, \ldots, G^{(T)} \rangle$ で表す．本章では，グラフ系列の中で $|V| = n$ の値は変化しないものとする．以下では重み付きグラフ系列を単にグラフ系列と呼ぶ．

図 2 はステップ数 $T = 4$ の重み付きグラフ系列の例である．同図では辺の重みが辺の太さで表されており，太い辺ほど大きな重み付けがされていることを表す．なお，簡略化のため重みが 0 の辺は図に示しておらず，以降のグラフでも重みが 0 の辺は図に描画されない．

時刻 t のグラフの頂点集合は，$\bigcup_{j=1}^{k} C_j^{(t)} = V$ を満たす k 個の互いに素な部分集合 $P^{(t)} = \left\{ C_1^{(t)}, C_2^{(t)}, \cdots, C_k^{(t)} \right\}$ に分割される．これらの表記を用いて，$1 \leq j \leq k$ に対してクラスタ系列は $C_j = \left\langle C_j^{(1)}, C_j^{(2)}, \cdots, C_j^{(T)} \right\rangle$ と表される．

本稿で扱う問題は，グラフ系列 $\langle G^{(1)}, G^{(2)}, \ldots, G^{(T)} \rangle$ と分割数 k が入力として与えられたとき，k 個のクラスタ系列 $\left\{ \left\langle C_j^{(1)}, C_j^{(2)}, \cdots, C_j^{(T)} \right\rangle \mid 1 \leq j \leq k \right\}$ を出力する問題である．この問題の応用例の 1 つは，1 章で述べた人間関係ネットワークの時系列から互いに結び付きの強いクラスタ（コミュニティ）を検出することである．人間関係ネットワークでは，連続する 2 時刻においてコミュニティが大きく変化することはほとんどないと考えられる．そこで，本稿が扱う問題では，以下の 2 つの要請を満たすものとする．

1. 各時刻のグラフでは，結び付きが強い頂点どうしが同一のクラスタに属する．
2. $C_j^{(t)}$ と $C_j^{(t+1)}$ でクラスタに属する頂点は大きく変わらない．これをクラスタの滑らかな変化と呼ぶ．

図 3 と**図 4** は図 2 からクラスタ系列を得た結果である．要請 2 を考慮しなければ，図 3 に示されるように，時刻 3 の辺 $(5, 6)$ は大きな重みを持つので，頂点 5 と 6 は同じクラスタに分割される．一方，要請 2 を考慮すると，時刻 3 の前後の分割と同様に頂点 6 は $C_2^{(3)}$ に分割され，$C_2^{(2)} = C_2^{(3)}$ となり，2 つのクラスタに属する頂点は変わらない．

3.2 Preserving Cluster Membership

グラフ $G = (V, E, w)$ の k 分割問題は，以下で定義される．これは RatioCut と呼ばれる関数の最少化問題である．

$$\min \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{e \in E(C_j, V \setminus C_j)} w(e)$$

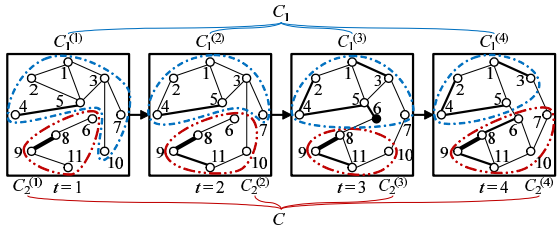

図 3 図 2 のグラフ系列と $k = 2$ に対する，クラスタ系列 (1)

Fig. 3 Cluster sequences (1) obtained from the graph sequence in Fig. 2.

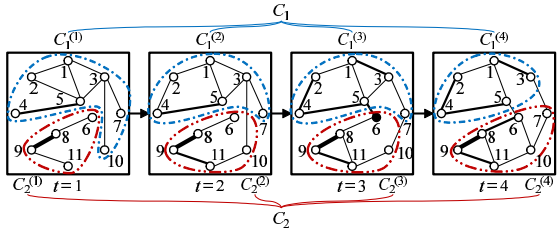

図 4 図 2 のグラフ系列と $k = 2$ に対する，クラスタ系列 (2)

Fig. 4 Cluster sequences (2) obtained from the graph sequence in Fig. 2.

ただし，$E(S, V \setminus S)$ は，一方の頂点が集合 S に含まれ，他方の頂点が集合 $V \setminus S$ に含まれる辺の集合である．また，スペクトラルクラスタリングを解説した文献 [16] では，上記の最少化問題は以下と等価であることが述べられている．

$$\min_{X \in \mathcal{R}^{n \times k}} \mathrm{Tr}(X^T L X) \qquad s.t.\ \ X^T X = I \ ^{*1} \tag{1}$$

ここで，X の (i, j) 要素 x_{ij} は以下の式で与えられる．

$$x_{ij} = \begin{cases} \frac{1}{\sqrt{|C_j|}} & \text{if } v_i \in C_j \\ 0 & \text{otherwise} \end{cases}$$

各頂点は，k 個のクラスタのうち 1 つのクラスタに属するので，I を n 次の単位行列とすると，$X^T X = I$ となる．また，L は G のラプラシアン行列であり，以下で定義される．行列の (i, j) 要素 a_{ij} が G の辺の重み $w((i, j))$ である G の隣接行列を A とする．また，$D = diag(\sum_{i=1}^{n} a_{i1}, \sum_{i=1}^{n} a_{i2}, \cdots, \sum_{i=1}^{n} a_{in})$ とする．このとき，$L = D - A$ である．

Preserving Cluster Membership（PCM）と呼ばれるグラフ系列のクラスタリング手法はオンライン型のアルゴリズムである．この手法では，グラフ $G^{(t)}$ が得られたとき，$G^{(t-1)}$ に対する X_{t-1} は既知である．PCM は以下を逐次的に最少化することにより，クラスタ系列を求める．

$$\min_{X_t \in \mathcal{R}^{n \times k}} \mathrm{Tr}(X_t^T L_t X_t) + \alpha |X_t X_t^T - X_{t-1} X_{t-1}^T|^2 \tag{2}$$

ただし，$\alpha \geq 0$ であり，L_t は $G^{(t)}$ のラプラシアン行列である．また，$X_t X_t^T$ の (i, j) 要素は，時刻 t において，i 番目の頂点と j 番目の頂点が同じクラスタに属しているとき正の実数となり，それ以外のとき 0 となる．また，行列の

*1 以下では $X^T X = I$ などの制約条件の表記を省略する．

絶対値の 2 乗は，行列の要素の 2 乗の総和である．式 (2) の目的関数の第 2 項が最少化されるということは，2 つ目の要請を満たすことを意味する．式 (2) の目的関数は以下のように変形できる．

$$\begin{aligned}
&\mathrm{Tr}(X_t^T L_t X_t) + \alpha |X_t X_t^T - X_{t-1} X_{t-1}^T|^2 \\
&= \mathrm{Tr}(X_t^T L_t X_t) + \alpha\ \mathrm{Tr}(X_t X_t^T X_t X_t^T \\
&\quad -2X_t X_t^T X_{t-1} X_{t-1}^T + X_{t-1} X_{t-1}^T X_{t-1} X_{t-1}^T) \\
&= \mathrm{Tr}(X_t^T L_t X_t) + 2\alpha k - 2\alpha\ \mathrm{Tr}(X_t^T X_{t-1} X_{t-1}^T X_t) \\
&= 2\alpha k + \mathrm{Tr}[X_t^T (L_t - 2\alpha X_{t-1} X_{t-1}^T) X_t]
\end{aligned}$$

したがって，式 (2) を最少化することは，以下を最少化することと同値である．

$$\min_{X_t \in \mathcal{R}^{n\times k}} \mathrm{Tr}\left[X_t^T \left(L_t - 2\alpha X_{t-1} X_{t-1}^T\right) X_t\right] \tag{3}$$

文献 [9] では，これを拡張したオフライン型アルゴリズムも示されている．オフライン型アルゴリズムを示すために，グラフ $G^{(t)}$ をクラスタリングするための最適化問題を示す．ここでは時刻 t の前後の時刻のクラスタリング結果 X_{t-1} と X_{t+1} が用いられる．

$$\min_{X_t \in \mathcal{R}^{n\times k}} \mathrm{Tr}\left[X_t^T \left(L_t - \alpha X_{t-1} X_{t-1}^T - \alpha X_{t+1} X_{t+1}^T\right) X_t\right] \tag{4}$$

式 (1) を最少化する X を求める関数を $func_1(L)$ とする．同様に，式 (3) と式 (4) を最少化する X_t を求める関数をそれぞれ $func_2(L_t, X_{t-1}, \alpha)$ と $func_3(L_t, X_{t-1}, X_{t+1}, \alpha)$ とする．これらを用いて，PCM オフライン型アルゴリズムを Algorithm 1 に示す．まず，$G^{(1)}$ をクラスタリングする．続いて，時刻 1 の結果を用いて $G^{(2)}$ をクラスタリングする．これを繰り返し，各時刻のグラフをクラスタリングする．その後，時刻 2 の結果を用いて $G^{(1)}$ をクラスタリングする．続いて，時刻 1 と 3 の結果を用いて $G^{(2)}$ をクラスタリングする．これを収束するまで繰り返す．

α を小さくすると，式 (2) の第 1 項を重視してクラスタリングされるため，要請 2 よりも要請 1 が重視され，各時刻のグラフが別々にクラスタリングされる．すなわち図 3 に示される結果となる．一方，α を大きくすると，式 (2) の第 2 項を重視してクラスタリングされるため，要請 1 よりも要請 2 が重視されて，滑らかな変化を検出でき，図 4 に示される結果となる．具体的には，図 3 の時刻 $t = 3$ のグラフにおいて頂点 6 はクラスタ系列 C_1 に属するが，その前後 $t = 2, 4$ では C_2 に属している．一方，図 4 では頂点 6 を C_1 にとどめたほうが式 (2) の第 2 項が小さくなるため，全時刻にわたって C_1 に属する．

3.3 PCM の課題

本節では，PCM オフライン型アルゴリズムの 3 つの課題をあげる．1 つ目の課題は PCM の性能は $G^{(1)}$ に依存する．$G^{(1)}$ に含まれる各クラスタの頂点が互いに結び付きが強くクラスリングが容易であれば，$G^{(2)}$ のクラスタリングでは X_1 が用いられるため，比較的容易になる．一方，$G^{(1)}$ で適切なクラスタリングが行われなければ，式 (2) の第 2 項によりその影響がそれ以後に波及する．

2 つ目の課題は PCM は各時刻でのクラスタの数が k であることに由来する．このため，時刻の経過とともに 1 つのクラスタが分裂し，直前の時刻よりもクラスタ数が増えるデータに適用しても適切なクラスタを検出できない．また，時刻の経過とともに 2 つのクラスタが併合し，直前の時刻よりもクラスタ数が減るデータに適用しても適切なクラスタを検出できない．時刻の経過とともに変化するグラフからクラスタ系列を検出するには，各時刻においてクラスタ数が k でない場合も許容すべきである．

3 つ目の課題は 2 つ目の課題と関連する．PCM では，各時刻の頂点数は n で，時刻が経過しもて一定であるとしている．しかし，ソーシャルネットワークの会員はつねに一定ではなく，時刻の経過とともに変化する．また，新たにネットワークに加わる会員だけでなく，そこから退会する者もいる．このようなデータにも適用可能なクラスタリング手法が必要となる．

Algorithm 1: PCM_offline

```
Data: ⟨G^(1), G^(2), ··· , G^(T)⟩, k
Result: X_1, X_2, ··· , X_T
for t ∈ [1, T] do
    if t = 1 then
        X_1 = func_1(L_1);
    else
        X_t = func_2(L_t, X_{t-1}, α);
while X_1, X_2, ··· , X_T が収束するまで do
    for t ∈ [1, T] do
        if t = 1 then
            X_1 = func_2(L_1, X_2, α);
        else
            if t = T then
                X_T = func_2(L_T, X_{T-1}, α);
            else
                X_t = func_3(L_t, X_{t-1}, X_{t+1}, α);
return X_1, X_2, ··· , X_T;
```

4. 提案手法

本章では，前章で述べた課題 1 と課題 2 を改善する手法を提案する．そのために以下を最少化する $X_1, X_2, \ldots, X_T \in \mathcal{R}^{n\times k}$ を求める問題を考える．

$$\sum_{t=1}^{T} \mathrm{Tr}(X_t^T L_t X_t) + \alpha' \sum_{t=1}^{T-1} |X_t - X_{t+1}|^2 \tag{5}$$

$$\begin{aligned}
&\mathrm{Tr}\left[X_1^T L_1 X_1\right] + \mathrm{Tr}\left[X_2^T L_2 X_2\right] + \alpha' \left|X_1 - X_2\right|^2 = \mathrm{Tr}\left[X_1^T L_1 X_1 + X_2^T L_2 X_2\right] + \alpha'\ \mathrm{Tr}[X_1^T X_1 + X_2^T X_2 - X_2^T X_1 - X_1^T X_2] \\
&= \mathrm{Tr}\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^T \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] + \alpha'\ \mathrm{Tr}\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^T \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \mathrm{Tr}\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^T \begin{pmatrix} L_1 + \alpha'\, I & -\alpha'\, I \\ -\alpha'\, I & L_2 + \alpha'\, I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] \\
&= \mathrm{Tr}\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^T \left\{ \begin{pmatrix} D_1 + \alpha'\, I & 0 \\ 0 & D_2 + \alpha'\, I \end{pmatrix} - \begin{pmatrix} W_1 & \alpha'\, I \\ \alpha'\, I & W_2 \end{pmatrix} \right\} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right]
\end{aligned}$$

図 5 2 ステップグラフ系列のクラスタリングの目的関数
Fig. 5 Objective function for a graph sequence with two steps.

$$\mathrm{Tr}\left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_T \end{pmatrix}^T \left\{ \underline{\begin{pmatrix} D_1 + \alpha' I & 0 & \cdots & \cdots & 0 \\ 0 & D_2 + 2\alpha' I & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & D_{T-1} + 2\alpha' I & 0 \\ 0 & \cdots & \cdots & 0 & D_T + \alpha' I \end{pmatrix}} - \underline{\underline{\begin{pmatrix} W_1 & \alpha' I & 0 & \cdots & 0 \\ \alpha' I & W_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & W_{T-1} & \alpha' I \\ 0 & \cdots & 0 & \alpha' I & W_T \end{pmatrix}}} \right\} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_T \end{pmatrix}\right]$$

図 6 T ステップグラフ系列のクラスタリングの目的関数
Fig. 6 Objective function for a graph sequence with T steps.

$$\mathrm{Tr}\left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_T \end{pmatrix}^T \left\{ \underline{\begin{pmatrix} D_1 + B_1 & 0 & \cdots & \cdots & 0 \\ 0 & D_2 + B_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & D_T + B_T \end{pmatrix}} - \begin{pmatrix} W_1 & \alpha' I & \alpha'\gamma I & \cdots & \alpha'\gamma^{T-2} I \\ \alpha' I & W_2 & \ddots & \ddots & \vdots \\ \alpha'\gamma I & \ddots & \ddots & \ddots & \alpha'\gamma I \\ \vdots & \ddots & \ddots & W_{T-1} & \alpha' I \\ \alpha'\gamma^{T-2} I & \cdots & \alpha'\gamma I & \alpha' I & W_T \end{pmatrix} \right\} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_T \end{pmatrix}\right]$$

図 7 忘却率を含んだ T ステップグラフ系列のクラスタリングの目的関数
Fig. 7 Objective function with forgetting rate for a graph sequence with T steps.

ここで，$\alpha' > 0$ である．式 (5) の第 1 項を最少化することは，各時刻のグラフ $G^{(t)}$ を要請 1 に従いクラスタリングすることに相当する．また，式 (5) の最少化が要請 2 を満たしていることを示すために，時刻 1 と 2 のみからなるグラフ系列を考える．このとき目的関数は

$$\mathrm{Tr}(X_1^T L_1 X_1) + \mathrm{Tr}(X_2^T L_2 X_2) + \alpha' |X_1 - X_2|^2$$

である．これより**図 5** に示される式が導出できる．同様にして，式 (5) より**図 6** に示される式を導出できる．図 6 の下線部の行列を D'，2 重下線部の行列を W' とし，$L' = D' - W'$ とすると，L' は以下を満たすグラフ G' のラプラシアン行列である．

- G' の頂点数は $n \times T$ である．これ以降，時刻 t における i 番目の頂点を $v_{t,i}$ で表す．
- $G^{(t)}$ に重みが $w((i,j))$ の辺 (i,j) が存在するならば，G' に重みが $w((i,j))$ の辺 $(v_{t,i}, v_{t,j})$ が存在する．
- 上記に加え，$1 \le t \le T-1, 1 \le i \le n$ に対して頂点 $v_{t,i}$ と $v_{t+1,i}$ の間には，重みが α' の辺が存在する．

したがって，式 (5) の最少化問題は G' の k 分割問題に帰着される．頂点 $v_{t,i}$ と $v_{t+1,i}$ の間には，重みが α' の辺が存在する．この辺がカットされるように，グラフ G' を k 分割すると，目的関数のとる値が大きくなるため，できるだけこの辺をカットしないように，すなわち $v_{t,i}$ と $v_{t+1,i}$ が同じクラスタに属するように分割がなされ，クラスタの滑らかな変化の検出が可能となる．以上より，要請 2 が満たされていることが分かる．**図 8** に 2 ステップからなるグラフ系列 $\langle G^{(1)}, G^{(2)} \rangle$ をグラフ G' に変換した例を示す．この図において，破線で描かれた辺の重みは α' である．

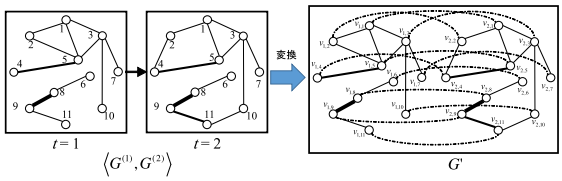

図 8 2 ステップからなるグラフ系列のグラフ G' への変換
Fig. 8 Conversion of a graph sequence with 2 steps into a graph G'.

式 (5) の最少化問題は G' の k 分割問題に帰着された．これにより，$G^{(1)}$ から逐次的にクラスタリングする PCM とは異なり，提案手法は G' を k 分割するので，得られる結果は $G^{(1)}$ のクラスタリングのしやすさに依存しない．これにより，PCM の持つ課題 1 を克服できている．さらに，あるクラスタには時刻 t 由来の頂点が含まれない可能性がある．すなわち，各時刻のグラフは k 個のクラスタに分割されるとは限らず，k 個以下のクラスタに分割される．このため，提案手法はある時刻 t でクラスタ数が k であるが，

Algorithm 2: 提案手法

Data: $\langle G^{(1)}, G^{(2)}, \cdots, G^{(T)} \rangle, k$
Result: $X_1, X_2, \cdots, X_T$
1 $\langle G^{(1)}, G^{(2)}, \cdots, G^{(T)} \rangle$ から G' を生成する．;
2 $\ell = n \times T$;
3 G' のラプラシアン行列 L' を求める（L' は $\ell \times \ell$ の行列）．;
4 L' の固有ベクトルを固有値が小さいもの順に k 個 $\vec{\gamma}_1, \vec{\gamma}_2, \cdots, \vec{\gamma}_k$ を求める．;
5 固有ベクトル $\vec{\gamma}_q$ を q 列目に持つ行列 $\Gamma \in R^{\ell \times k}$ を生成する．;
6 Γ について i 番目の行を k 次元空間上の点として扱い，ℓ 個の点を k-means を用いてクラスタリングする．その結果を $P_1, P_2, \cdots, P_k$ とする．;
7 **for** $t \in [1, T]$ **do**
8 　$X_t = 0$;
9 **for** $j \in [1, k]$ **do**
10 　**for** $v_{t,i} \in P_j$ **do**
11 　　$x_{i,j}^{(t)} = \frac{1}{\sqrt{|\{v_{t',i'} \in P_j | t = t'\}|}}$;
12 **return** $X_1, X_2, \cdots, X_T$;

時刻 $t+1$ においてクラスタ数が $k-1$ であるようなクラスタ系列の出力が可能となる．したがって，提案手法は先述の課題 2 を克服できる．

式 (5) の次行において，$\alpha' \geq 0$ と書かれていないことに注意されたい．$\alpha' = 0$ であれば，G' において $G^{(t)}$ と $G^{(t+1)}$ の間には辺が存在しない．このため，G' を，たとえば T 分割すると，各時刻のグラフがクラスタになる．そのため，α' は正の実数である必要がある．また，PCM の目的関数と提案手法の目的関数は非常に似ている．しかし，

$$\mathrm{Tr}(X_1^T L_1 X_1) + \mathrm{Tr}(X_2^T L_2 X_2) + \alpha |X_1^T X_1 - X_2^T X_2|^2$$

では，図 5 のような導出ができない．つまり，PCM の目的関数の最少化問題を，あるグラフの k 分割問題には帰着できない．

提案手法の擬似コードを Algorithm 2 に示す．提案手法は 3–6 行目で文献 [16] のスペクトラルクラスタリングのアルゴリズムを含んでいる．7–8 行目で，X_t をゼロ行列で初期化し，9–11 行目では，頂点 $v_{t,i}$ が j 番目のクラスタ P_j に含まれているならば，行列 X_t の (i,j) 要素 $x_{i,j}^{(t)}$ に $\frac{1}{\sqrt{|\{v_{t',i'} \in P_j | t = t'\}|}}$ を代入する．

Algorithm 2 に従うと，各時刻での頂点数は必ずしも n である必要はない．2 行目の $\ell = n \times T$ を $\ell = \sum_{t=1}^{T} |V(G^{(t)})|$ に置き換えることで，先述の課題 3 を克服することができる．ここで $|V(G^{(t)})|$ は時刻 t のグラフの頂点数である．

5. 忘却率を含めたグラフ系列のクラスタリング

前章では連続する 2 時刻間でのクラスタの滑らかな変化を考慮した手法を提案した．本章では，その考えをさらに拡張し，τ 離れた時刻間でのクラスタの滑らかな変化を考慮した手法を考える．これを以下で定式化する．

$$\sum_{t=1}^{T} \mathrm{Tr}(X_t^T L_t X_t) + \alpha' \sum_{\tau=1}^{T-1} \gamma^{\tau-1} \sum_{t=1}^{T-\tau} |X_t - X_{t+\tau}|^2 \quad (6)$$

ここで $0 \leq \gamma \leq 1$ であり，γ を忘却率と呼ぶ．また，$\gamma = 0$ のとき式 (5) と一致するので，式 (6) は式 (5) の一般形である．前章と同様に式 (6) を整理すると，**図 7** に示される式が得られる．ここで $B_t = \alpha' \sum_{\tau=1, \tau \neq t}^{T} \gamma^{|t-\tau|-1} I$ であり，対角行列である．図 7 の式の下線部は以下を満たすグラフ G'' のラプラシアン行列 L'' である．

- G'' の頂点数は $n \times T$ である．
- $G^{(t)}$ に重みが $w((i,j))$ の辺 (i,j) が存在するならば，G' に重みが $w((i,j))$ の辺 $(v_{t,i}, v_{t,j})$ が存在する．
- 上記に加え，$1 \leq t < t' \leq T$，$1 \leq i \leq n$ に対して頂点 $v_{t,i}$ と $v_{t',i}$ の間には，重みが $\alpha' \gamma^{(t'-t-1)}$ の辺が存在する．

G' は G'' の部分グラフであり，$\gamma = 0$ のとき，G' と G'' は同型になる．また，G'' の場合でも，Algorithm 2 の G' を G'' に，L' を L'' に変更するだけで，Algorithm 2 を用いることができる．前章で，提案手法が PCM の課題 3 を克服できることを述べた．しかし，$G^{(t)}$ に頂点 v_i が存在し，その前後の時刻である $G^{(t-1)}$ と $G^{(t+1)}$ の両方で，頂点 v_i が存在しないとき，この頂点に対するクラスタの滑らかな変化を考慮することができなかった．しかし，本章で導入した忘却率に対して減衰する重みを持つ辺を用いることで，我々の手法は PCM の課題 3 をさらに克服できる．

6. 評価実験

6.1 実験設定

本章では，PCM と提案手法の比較実験を行う．評価には Adjusted Rand Index（ARI）を用いる．ARI とはデータ生成時のクラスタをどの程度再現できるかを計る指標であり，比較するクラスタに共通して含まれる頂点数を用いて次のように計算される．n 個の頂点集合が互いに素な r 個の部分集合 $\mathcal{U} = \{U_1, \cdots, U_r\}$ と c 個の部分集合 $\mathcal{V} = \{V_1, \cdots, V_c\}$ に分割されているとする．ただし，$\sum_{i=1}^{r} |U_i| = \sum_{j=1}^{c} |V_j| = n$ である．次に n_{ij} を U_i と V_j に共通して含まれる頂点数 $|U_i \cap V_j|$ とすると，n_{ij} はクラスタの組合せによって**表 1** に示すとおりになる．$n_{i\cdot}$ はクラスタ U_i の頂点数，$n_{\cdot j}$ はクラスタ V_j の頂点数を表す．このとき，U_i と V_j に共通して含まれる頂点数の個数は $\binom{n_{ij}}{2}$ で計算される．したがって，ARI は次のように計算され，ARI は 0 以上 1 以下の値をとる．

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}$$

実験結果では分割 $\mathcal{U}$ は生成時に設定した本来の分割に，$\mathcal{V}$ はクラスタリングを適用して得られた分割に対応するため，

表 1 分割 $\mathcal{U}$, $\mathcal{V}$ を比較する分割表
Table 1 Contingency table for partitions $\mathcal{U}$ and $\mathcal{V}$.

$\mathcal{U} \setminus \mathcal{V}$	V_1	V_2	$\cdots$	V_c	合計
U_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
U_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
U_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
合計	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	$n_{..} = n$

表 2 人工データ生成のパラメータ
Table 2 Parameters of the artificial data.

パラメータ	既定値
頂点数 n	1,100 (=600+300+200)
クラスタ系列数 k	3
半径 r	3
分散 var	1.0
ステップ数 T	10
振幅 A	1.0
初期位相 φ	0
周期 ω	$\frac{\pi}{4}$
移動する頂点数 m	5
近傍数 κ	10

ARI が 1 に近づくほど良い結果が得られたことを表す．また，本稿で示す ARI は 20 回試行した結果の平均である．

6.2 実験結果

6.2.1 課題 1 克服の検証実験

PCM と提案手法の比較実験のために，以下の手順に従い人工的に実験データの生成する．2 次元平面上の原点を中心とする半径 r の円周上に k 個の点を等間隔に配置し，これらの点の座標を k 個のガウス分布の平均とする．そして，各平均に対して，分散 var のガウス分布に従う点集合を生成する．この実験での 3 つの点集合内の点の数はそれぞれ 600，300，200 とする．生成されたそれぞれの点集合が 1 つのクラスタに相当する．ただし，時刻を経るごとに各ガウス分布の平均が原点に近づいたり離れたりするようにして，点集合を生成する．そこで，ガウス分布の平均は初期位相 φ，振幅 A で，半径 r の円周上を中心に周期 ω で正弦的に振動させて，ガウス分布に従う点集合を生成する．すなわち，時刻 t で i 番目（$1 \leq i \leq k$）のクラスタに相当するガウス分布の平均座標 (x, y) は，$R(\theta)$ を回転角 θ の回転行列として，次のとおりである．

$$\begin{pmatrix} x \\ y \end{pmatrix} = R\left(\frac{2\pi(i-1)}{k}\right)\left[A\begin{pmatrix} \sin(\omega(t-1)+\varphi) \\ 0 \end{pmatrix} + \begin{pmatrix} r \\ 0 \end{pmatrix}\right] \tag{7}$$

このままでは，α' に大きな値を設定することで高い ARI を得ることができるため，過大に評価されることになる．これを避けるために，各時刻において，最もの大きい点集合からランダムに m 個を選び，2 番目に大きい点集合に属する点となるように点を移動させる．以上で生成された各時刻の点集合をグラフの頂点集合，2 頂点間のユークリッド距離を L とするとき，辺の重みを $\exp\left(-\frac{L^2}{2}\right)$ とする各時刻のグラフを κ 最近傍グラフ [16] により生成した．さらに，ステップ数が T のグラフ系列を生成した．

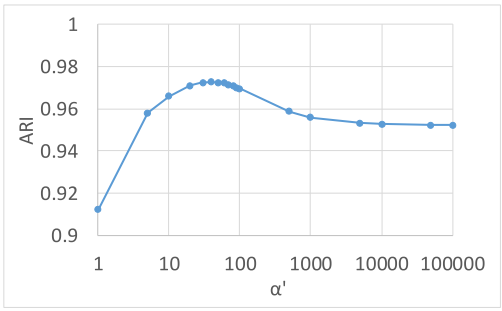

図 9 α' を変化させたときの ARI
Fig. 9 ARIs for the proposed method for various values of α'.

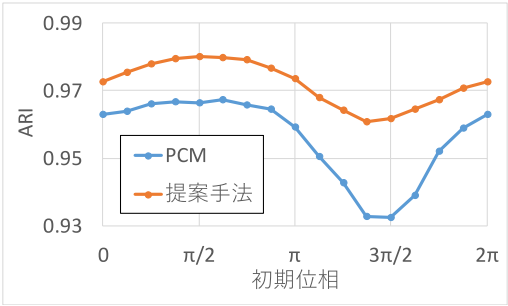

図 10 φ を変化させたときの ARI
Fig. 10 ARIs for various values of φ.

図 9 は，提案手法において α' を 1 から 100,000 に変化させたときの ARI の平均値である．$\alpha' = 40$ のとき要請 1 と要請 2 のバランスがとれ，ARI は最大となる．これよりも α' を大きくすると，式 (5) の第 2 項の影響が強くなり，要請 2 を満たそうとするので，ARI の値は減少する．一方，α' を小さくすると，式 (5) の第 1 項の影響が強くなり，要請 1 を満たそうとするので，ARI の値は減少する．さらに，α' を 1 よりも小さくすると，各時刻のグラフの頂点を結ぶ辺の重みよりも，連続する 2 時刻間の間の辺の重み α' が小さくなり，G' の辺のうち重み α' の辺をカットして，クラスタを構成しようとするので，ARI は大幅に低下する．以上より α' を調整することにより，要請 1 と要請 2 を満たすクラスタ系列を得ることが確認できた．これ以降，α' の値を 40 に設定する．PCM に関しても，同様の結果が得られたので，α の値を 4 に設定する．

図 10 は，φ を 0 から 2π まで変化させたときの実験結果である．PCM の ARI は $\varphi = \frac{3}{2}\pi$ の周辺で大きく低下する．$\varphi = \frac{3}{2}\pi$ のとき，$t = 1, 5, 9$ において，クラスタの平均が互いに近づくため，互いのクラスタの点の分布は 3 度大きく重なる．特に，時刻 1 のグラフ $G^{(1)}$ で適切なクラスタ

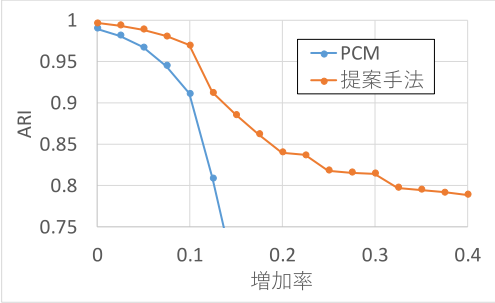

図 11 β を変化させたときの ARI
Fig. 11 ARIs for various values of β.

リングが行われず，式 (2) の第 2 項によりその影響がそれ以後の時刻に波及しているため，ARI は大きく低下する．一方，提案手法の ARI は $\varphi = \frac{3}{2}\pi$ の周辺で低下するものの，いかなる φ において PCM の ARI を上回る．これにより，提案手法は PCM の持つ課題 1 を解決できていることが分かる．

6.2.2 課題 2 克服の検証実験

本実験で，3 クラスタのうち，頂点数が少ない 2 つのクラスタの平均は式 (7) ではなく，式 (8) に従うように人工データを作成した．

$$\begin{pmatrix} x \\ y \end{pmatrix} = R\left(\frac{2\pi(i-1)}{k}\right) r \begin{pmatrix} \exp\left[\beta(t-T)\right] \\ 0 \end{pmatrix} \qquad (8)$$

つまり，2 つのクラスタの平均は時刻 1 において，互いに近い位置にあるが，時刻とともに指数関数的に離れていき，時刻 T において，原点を中心とする半径 r の円周上にのる．また，1 つのクラスタの平均と別のクラスタの平均との距離が $2var$ より大きい場合，2 つの平均から生成される頂点は異なるクラスタに属すると見なすことで，1 つクラスタが 2 つのクラスタに分裂したことにする．

図 11 は，増加率 β を 0 から 0.4 に変化させたときの実験結果である．β が小さいとき，$G^{(1)}$ において，3 つのクラスタは互いに離れており，その後もほとんど近づくことはないので，PCM の ARI は大きい．しかし，β を大きくすると，時刻が小さいとき，検出すべきクラスタ数は 2 であるのに対して，PCM は RatioCut に基づいて頂点数のバランスのとれた 3 つのクラスタに分割しようとするので，ARI は小さくなる．一方，提案手法は，グラフ系列を 1 つのグラフ G' に変換し，G' の k 分割問題を解く．このため，各時刻が k 個のクラスタとなるとは限らず，k 個以下のクラスタになる．よって，様々な β に対して適切なクラスタを検出できる．

図 12 は，式 (7) と式 (8) から得られた点の集合の散布図であり，提案手法で得られたクラスタ系列に従って色分けされている．この結果は，各時刻の頂点数 $n = 110$，増加率 $\beta = 0.35$ のときに得られた結果である．クラスタ 1 は正弦的に振動するクラスタなので，時刻が経過しても頂点数はほぼ一定である．一方，クラスタ 3 は時刻 1 において頂点数は 0 であったが，時刻の経過とともに頂点数が増え，時刻 10 で 18 となった[*2]．これにより，1 つのクラスタが 2 つに分裂する場合に，提案手法は適切なクラスタ系列を出力できる．

本実験では，1 つのクラスタが 2 つに分裂する人工データを作成して実験を行った．提案手法は，グラフ系列を 1 つのグラフ G' に変換し，k 分割問題を解くので，時間の向きには依存しない．式 (8) の代わりに式 (9) を用いると，クラスタが併合する人工データを生成することができるが，このデータを用いても提案手法の結果は図 11 と同様となる．

$$\begin{pmatrix} x \\ y \end{pmatrix} = R\left(\frac{2\pi(i-1)}{k}\right) r \begin{pmatrix} \exp\left[-\beta(t-1)\right] \\ 0 \end{pmatrix} \qquad (9)$$

以上により，PCM の課題 2 を克服できていることが分かる．

6.2.3 課題 3 克服の検証実験

図 13 は，6.2.1 項と同様の方法で人工データを作成し，$ratio$% の確率で頂点を除去した人工データに対する実験結果である．この実験では，忘却率 γ を 0 から 0.5 に変更して，ARI を計測した．$ratio = 0$ でかつ $\gamma = 0$ のとき，6.2.1 項の実験設定と同じである．また，このデータに対して PCM は適用できないため，PCM の実験結果は掲載されていない．γ を大きくすると，Δ 時刻離れた頂点が $\alpha'\gamma^{\Delta-1}$ の重みを持つ辺でつながるので，要請 2 を満たすことができ，ARI は上昇する．特に，γ を 0 から 0.025 に増やしたときにその効果は最も大きい．さらに，γ の値を増加させると，要請 2 の影響が強くなりすぎるため，ARI が減少する．この傾向は，6.2.1 項の結果と一致する．ただし，$ratio = 0$ のときは，6.2.1 項の実験において α' が十分に調整されているので，γ を大きくすると，要請 2 の影響が強くなり ARI は減少する．以上により，PCM の持つ課題 3 を克服できていることが分かる．なお，$\gamma \geq 0.1$ のときに $ratio$ が大きいほうが ARI の値が大きくなる理由は，以下のとおりである．$ratio = 0$ で $\gamma = 0$ のときのグラフ $G^{(t)}$ は，約 $\kappa|V|/2$ 個の辺を持ち，$G^{(t+1)}$ と $|V|$ 個の辺でつながるため，前者のほうのつながりが強い．これに対して，$ratio$ や γ の値を大きくすると，上記の 2 種類の辺のバランスがとれるため，同じ γ の値に対して，$ratio$ が大きいほうが ARI の値が大きくなる．

7. 考察

文献 [11] では，近年さかんに研究されているグラフ系列のタイプがまとめられている．解析アルゴリズムの設計のしやすさから，時刻が経過しても頂点数が変化しないグラ

[*2] 検出された 3 番目のクラスタ系列の頂点数は 0，1，4，8，13，16，16，18，18，18 と時刻の経過ともに増加した．

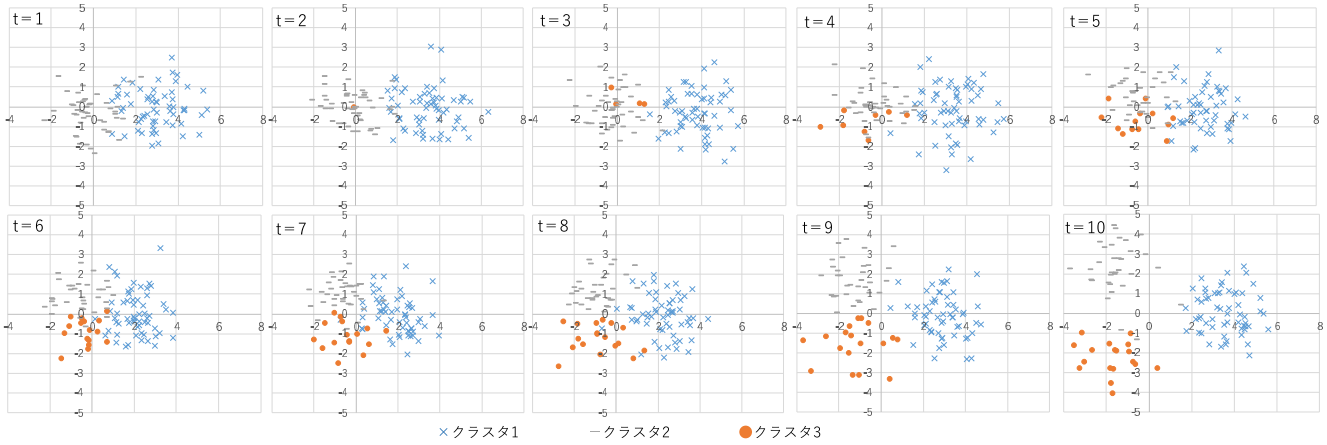

図 12 検出されたクラスタ系列の頂点の分布
Fig. 12 Distribution of vertices in detected cluster sequences.

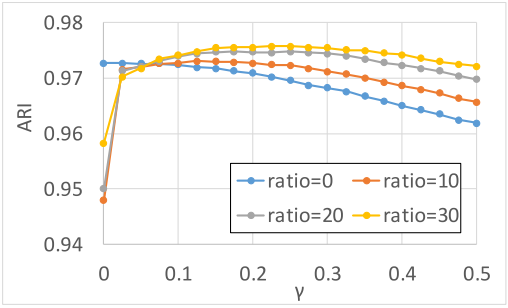

図 13 *ratio* と γ を変化させたときの提案手法の ARI
Fig. 13 ARIs for various values of *ratio* and γ.

フ系列のクラス，あるいは時間の経過とともに頂点が単調に増加するグラフ系列のクラスが扱われることが多いが，本稿で対象としたグラフ系列は最も汎用的なグラフ系列のクラスであり，グラフ系列の頂点数は，時間の経過とともに増加することも，減少することもある．また，本稿で提案したアルゴリズムは，このような汎用的なグラフ系列に対して，適用可能なアルゴリズムである．

一般にクラスタリングの問題において，ハイパーパラメータである分割数 k は試行錯誤を繰り返して，データに合致する適切な値に調整することが多い．提案手法で用いられる α' や γ も同様である．これらの値の決め方は以下のとおりである．α' と γ に適当な値を設定したうえで，提案手法を実行する．得られた結果の $C_j^{(t)}$ と $C_j^{(t+1)}$ を比べ，様々な j や t に対して $C_j^{(t)}$ と $C_j^{(t+1)}$ が一致するようであれば，要請 2 が強すぎるので，α' や γ を小さくする．一方，$C_j^{(t)}$ と $C_j^{(t+1)}$ が大きく異なり，クラスタの滑らかな変化が得られていないようであれば，要請 1 が強すぎるので，α' や γ を大きくし，試行錯誤を繰り返し，α' や γ を調整する．

提案手法は，その内部でスペクトラルクラスタリングを用いる．スペクトラルクラスタリングは離散問題である RatioCut の連続緩和問題を解くので，外れ値に頑健であるという特徴を持ち，提案手法もこの特徴を継承している．

本稿の式 (5) の代わりに

$$\min \sum_{t=1}^{T} \sum_{j=1}^{k} \sum_{e \in E(C_j^{(t)}, V \setminus C_j^{(t)})} w^{(t)}(e) + \alpha' \sum_{t=1}^{T+1} |X_t - X_{t+1}| \tag{10}$$

とし，X_t を

$$x_{ij}^{(t)} = \begin{cases} 1 & \text{if } v_i \in C_j^{(t)} \\ 0 & \text{otherwise} \end{cases}$$

と 0/1 行列とすると，$|X_t - X_{t+1}|/2$ はクラスタ間で移動する頂点数となるので，クラスタの大きさ $|C_j^{(t)}|$ に依存せず要請 2 に関するペナルティを掛けることができるという利点を持つ．しかし，この式から図 6 のような式を導出することは困難であるので，スペクトラルクラスタリングに帰着することができない．そのため，式 (10) の最適化問題では外れ値に弱くなる．

PCM，および提案手法で最も計算時間を要するのが，ラプラシアン行列から固有ベクトルを得るための計算である．行列の次数を n とすると，固有値計算は $O(n^3)$ である．PCM は逐次的に $G^{(t)}$ をクラスタリングするため，その計算量は $O(Tn^3)$ である．一方，提案手法は，グラフ系列を 1 つのグラフ G' に変換したうえで，G' をクラスタリングするため，その計算量は $O(T^3n^3)$ である．このため，グラフ系列の長さが大きくなったり，グラフの頂点数が増えたりすると，提案手法の計算時間は非常に大きくなる．これについては，今後の課題となる．

8. まとめ

本稿では，グラフ系列を対象として，クラスタの滑らかな変化を考慮したクラスタリング手法を提案した．提案手法はスペクトラルクラスタリングに基づくため外れ値に対して頑健であり，PCM と呼ばれる従来手法の持つ課題を克服している．具体的には，提案手法により得られる結果

は初期のグラフである $G^{(1)}$ のクラスリングのしやすさに依存しない，各時刻のグラフ $G^{(t)}$ をちょうど k ではなく k 以下のクラスタに分割するので，時間の経過とともに分裂，あるいは併合していくクラスタを検出することができる．また，各時刻のグラフの頂点数が一定である必要もないので，会員が新たに入会したり，退会したりするソーシャルネットワークに適用して，そこからクラスタ（コミュニティ）の変化を検出することが可能である．ただ，このような実データに適用するためには，計算時間に関する課題を持つので，今後，その課題に取り組む予定である．

謝辞 本研究の立ち上げの時期に，貴重な助言をいただいた大阪大学の鷲尾隆教授と河原吉伸准教授に感謝の意を表する．

参考文献

[1] Aggarwal, C.C., Han, J., Wang, J. and Yu, P.: A framework for clustering evolving data streams, *Proc. International Conference on Very Large Data Bases*, pp.81–92 (2003).

[2] Aggarwal, C.C., Han, J., Wang, J. and Yu, P.: A framework for projected clustering of high dimensional data streams, *Proc. International Conference on Very Large Data Bases*, pp.852–863 (2004).

[3] Aggarwal, C.C., Han, J., Wang, J. and Yu, P.: On demand classification of data streams, *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, pp.503–508 (2004).

[4] Bar-Joseph, Z., Gerber, G., GiRord, D.K., Jaakkola, T.S. and Simon, I.: A new approach to analyzing gene expression time series data, *Proc. International Conference on Computational Biology*, pp.39–48 (2002).

[5] Beringer, J. and Aullermeier, E.: Online clustering of parallel data streams, *Data and Knowledge Engineering*, pp.180–204 (2006).

[6] Cao, F., Estery, M., Qian, W. and Zhou, A.: Density-based clustering over an evolving data stream with noise, *Proc. SIAM Conference on Data Mining*, pp.328–339 (2006).

[7] Chakrabarti, D., Kumar, R. and Tomkins, A.: Evolutionary clustering, *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, pp.554–560 (2006).

[8] Charikar, M., O'Callaghan, L. and Panigrahy, R.: Better streaming algorithms for clustering problems, *Proc. ACM Symposium on Theory of Computing*, pp.30–39 (2003).

[9] Chi, Y., Song, X., Zhou, D., Hino, K. and Tseng, B.L.: On Evolutionary Spectral Clustering, *ACM Trans. Knowledge Discovery from Data*, Vol.3, No.4 (2009).

[10] Domingos, P. and Hulten, G.: A general method for scaling up machine learning algorithms and its application to clustering, *Proc. International Conference on Machine Learning*, pp.106–113 (2001).

[11] Inokuchi, A. and Washio, T.: Mining Frequent Graph Sequence Patterns Induced by Vertices, *Proc. SIAM International Conference on Data Mining*, pp.466–477 (2010).

[12] 小西哲生，猪口明博：スペクトラルクラスタリングに基づいた動的に変化するグラフのクラスタリング，第 76 回情報処理学会全国大会（IPSJ），2L-2，講演論文集，pp.1-335–1-336 (2014).

[13] Möller-Levet, C.S., Klawonn, F., Cho, K.-H., Yin, H. and Wolkenhauer, O.: Clustering of unevenly sampled gene expression time-series data, *Fuzzy Sets and Systems*, Vol.152, No.1, pp.49–66 (2005).

[14] O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S. and Motwani, R.: Streaming-data algorithms for high-quality clustering, *Proc. IEEE International Conference on Data Engineering*, pp.685–694 (2002).

[15] van Wijk, J.J. and van Selow, E.R.: Cluster and calendar based visualization of time series data, *IEEE Symposium on Information Visualization*, pp.4–9 (1999).

[16] von Luxburg, U.: A tutorial on spectral clustering, *Statistics and Computing*, Vol.17, No.4, pp.395–416 (2007).

[17] Wang, Y., Liu, S., Feng, J. and Zhou, L.: Mining naturally smooth evolution of clusters from dynamic data, *Proc. SIAM International Conference on Data Mining*, pp.125–134 (2007).

長村 佳歩

1993 年生．2016 年関西学院大学理工学部情報科学科卒業．奈良先端科学技術大学院大学情報科学研究科博士前期課程在学．機械学習，音声翻訳の研究に興味を持つ．

奥井 颯平

1992 年生．2015 年関西学院大学理工学部情報科学科卒業．同大学大学院博士前期課程在学．情報検索，データマイニング，機械学習の研究に興味を持つ．

猪口 明博 （正会員）

1975 年生．2000 年大阪大学大学院工学研究科通信工学専攻博士前期課程修了．同年日本アイ・ビー・エム（株）に入社，東京基礎研究所に配属．2004 年大阪大学大学院工学研究科通信工学専攻博士後期課程修了．博士（工学）．大阪大学産業科学研究所助教を経て，現在に至る．データマイニング，機械学習，テキストマイニング，データベースの研究・開発に従事．2002 年 Journal of Computer Aided Chemistry 論文賞．