2L-2

スペクト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

小西 哲生 †　　猪口 明博 ‡
†‡ 関西学院大学 理工学部 情報科学科

1 はじめに

人間関係ネットワークは人を頂点，人間関係を辺とするグラフで表現することができる．また，時間とともに変化する人間関係ネットワークは動的に変化するグラフで表すことができる．グラフにおいて，互いに結びつき強い頂点の集合をクラスタとすると，動的に変化するグラフのクラスタを発見することでクラスタの変化を検出することができる．本稿では，スペクトルクラスタリングに基づく動的グラフのクラスタリング手法を提案し，その性能を評価する．

2 グラフ系列のクラスタリング

本節では文献 [2] で提案されたグラフ系列のクラスタリング手法を紹介する．時刻 t のグラフを $G^{(t)} = \left(V, E, w^{(t)}\right)$ で表す．また，E は全ての辺の集合 $E = V \times V$ 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 $\left\{G^{(1)}, \ldots, G^{(T)}\right\}$ で表す．本研究では，グラフ系列の中で $|V|$ の値は変化しないものとする．図 1 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 の辺は図に示されない．

次に，グラフ系列内の各グラフに含まれる頂点を 1 以上最大 k 個のクラスタに分割し，グラフ系列をクラスタ系列に分割するためのコスト関数を説明する．まず，クラスタ系列を定義する．グラフ $G^{(t)}$ に含まれる頂点を，k 分割して得られるクラスタを $P^{(t)} = \left\{C_1^{(t)}, \cdots, C_k^{(t)}\right\}$ とする．この時，クラスタ $C_i^{(t)}$ が空集合であることを許容する．また，$P = \left\{P^{(1)}, \cdots, P^{(T)}\right\}$ とする．さらに，時刻 t におけるクラスタと時刻 $t+1$ におけるクラスタを対応付ける $M = \left\{M^{(1)}, \cdots, M^{(T-1)}\right\}$ を，$1 \le t \le T-1$, $1 \le i^{(t)} \le k$, 及び $1 \le j^{(t)} \le k$ を満たす $i^{(t)}$ と $j^{(t+1)}$ に対して以下のように定義し，$\left(i^{(t)}, j^{(t+1)}\right) \in M^{(t)}$ であるとき，クラスタ $C_{i^{(t)}}^{(t)}$ とクラスタ $C_{j^{(t+1)}}^{(t+1)}$ は対応すると呼ぶ．

- $\left|\left\{\left(i^{(t)}, j^{(t+1)}\right) \in M^{(t)} \mid C_{j^{(t+1)}}^{(t+1)} \in P^{(t+1)}\right\}\right| = 1$ for $C_{i^{(t)}}^{(t)} \in P^{(t)}$
- $\left|\left\{\left(i^{(t)}, j^{(t+1)}\right) \in M^{(t)} \mid C_{i^{(t)}}^{(t)} \in P^{(t)}\right\}\right| = 1$ for $C_{j^{(t+1)}}^{(t+1)} \in P^{(t+1)}$
- $|M^{(t)}| = k$

Clustering a Graph Sequence using a Spectral Clustering
†Akio KONISHI　‡Akihiro Inokuchi
†Department of Informatics, School of Science and Technology, Kwansei Gakuin University

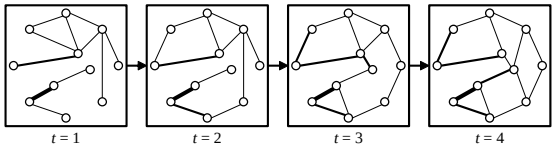

図 1: 4 ステップ重み付きグラフ系列の例

以上より，クラスタ系列を以下のように定義する．

定義 2.1 (クラスタ系列) P と M が与えられた時，$j^{(2)} = i^{(2)}, \cdots, j^{(T-1)} = i^{(T-1)}$ を満たす $\left(i^{(1)}, j^{(2)}\right) \in M^{(1)}, \cdots, \left(i^{(T-1)}, j^{(T)}\right) \in M^{(T-1)}$ に対して，クラスタ系列 C_h は，$C_{i^{(1)}}^{(1)} \in C_h, \cdots, C_{i^{(T)}}^{(T)} \in C_h$ を満たす．■

続いて，コスト関数の定義に用いられる，隣接する時刻における 2 つのクラスタ間の距離を定義する．

定義 2.2 (クラスタ間の距離) 時刻 t $(1 \le t \le T)$ のグラフ $G^{(t)} = \left(V, E, w^{(t)}\right)$ を分割したクラスタ $C_i^{(t)}$ に対して，$C_i^{(t)}$ に含まれる頂点を表すベクトルを $A_i^{(t)} = \left(a_1^{(t)} \ldots a_{|V|}^{(t)}\right)$ とする．ただし，

$$a_l^{(t)} = \begin{cases} 1 & if\ \ v_l \in C_i^{(t)} \\ 0 & otherwise \end{cases} \qquad (1 \le l \le |V|) \tag{1}$$

である．この時，$C_i^{(t)}$ と $C_j^{(t+1)}$ の距離を次式で定義する．

$$dist\left(C_i^{(t)}, C_j^{(t+1)}\right) = \sum_{l=1}^{|V|} \left|a_l^{(t+1)} - a_l^{(t)}\right| \blacksquare \tag{2}$$

次に，グラフ系列の分割の良さを定量的に評価するコスト関数を，2 つの観点から示す．本稿では，次の条件を満たすクラスタ系列を良いクラスタ系列と考える．

1. 各時刻のグラフでは，結び付きが強い頂点同士が同一のクラスタに属す．
2. クラスタ系列の全時刻にわたる，対応するクラスタ間の距離の和が最小である．

$G^{(t)}$ に含まれる頂点を k 個のクラスタ $C_i^{(t)}$ に分割するためのコストを，カット関数を用いて $\sum_{i=1}^{k} cut\left(C_i^{(t)}\right)$ とする．さらに，この $G^{(t)}$ の分割コストの全時刻にわたる和を $F_1(P)$ として，次のように定義する．

$$F_1(P) = \sum_{t=1}^{T}\sum_{i=1}^{k} cut\left(C_i^{(t)}\right) = \sum_{t=1}^{T}\sum_{i=1}^{k}\sum_{e \in E(C_i^{(t)}, V \setminus C_i^{(t)})} w^{(t)}(e)$$

カット関数の値はクラスタを作るために切られる辺の重みの和である．従って，関数 $F_1(P)$ の値が小さくなる P に分割することで，頂点間の重みが大きい頂点同士を同一のクラスタにまとめることができる．

クラスタ系列内の対応するクラスタ間の距離に基づいた，コスト関数を $F_2(P, M)$ とし，次式で定義する．

$$F_2(P, M) = \sum_{t=1}^{T-1}\sum_{(i,j) \in M^{(t)}} dist\left(C_i^{(t)}, C_j^{(t+1)}\right) \tag{3}$$

$F_2(P,M)$ は全ての隣接する時刻における，対応するクラスタ間の距離の和を表し，この関数の値が小さくなる分割 P と M を求めることで，良いクラスタ系列の条件 2 番を満たすクラスタ系列を構成することができる．

先に述べた 2 つのコスト関数をまとめて，グラフ系列を 1 以上最大 k 個のクラスタ系列に分割するコスト関数 $F(P,M)$ を次のように定義する．

$$F(P,M) = F_1(P) + \alpha F_2(P,M) \quad (\alpha > 0) \tag{4}$$

以上より，クラスタの変化を発見するためにグラフ系列のクラスタリング問題を次のように定義する．

問題 2.3 (グラフ系列のクラスタリング問題) グラフ系列とクラスタ系列数 k が入力として与えられた時，コスト関数 $F(P,M)$ が最小となる分割 P と M を出力する．

文献 [2] では，コスト関数 F が劣モジュラ関数の和であることが示されている．コスト関数 F は劣モジュラ関数の和であるので，劣モジュラ最小化問題を利用して解くことができる．しかしながら，集合 S に対して劣モジュラ関数が定義され，その劣モジュラ最小化問題を Queyranne のアルゴリズム [3] で解くと，その計算量は $O(|S|^3)$ となるため大規模グラフ系列に適用できない．また，最適解を求めるこの手法では外れ値の存在するデータに対してクラスタリングの制度が悪くなる場合がある．この課題に対処するために，本稿ではコスト関数がカット関数であることを示し，カット関数最小化問題に帰着して解く手法を提案する．

3 提案手法

グラフ系列 $\{G^{(1)},\cdots,G^{(T)}\}$ に対して，nT 個の頂点からなる無向グラフ $G' = (V', E', w')$ を考える．V' の頂点を $v_{t,i}$ で表し，$v_{t,i}$ はグラフ系列の t 番目のステップにおける i 番目の頂点に相当するものとする．また，G' における重み関数 w' を $e \in E' = V' \times V'$ に対して以下のように定義する．

$$w'(e(v_{t,i}, v_{t',j})) = \begin{cases} w^{(t)}(e(v_i, v_j)) & if \ \ t = t' \\ \alpha & if \ \ t' = t+1, i = j \\ 0 & otherwise \end{cases} \tag{5}$$

すなわち，$v_{t,i}$ と $v_{t',j}$ が同じ時刻ステップのグラフ $G^{(t)}$ 由来の頂点なら，$v_{t,i}$ と $v_{t',j}$ の間の辺の重みは $G^{(t)}$ における対応する辺の重みとする．また，$v_{t,i}$ と $v_{t',j}$ が 1 時刻ステップの異なるグラフ由来の頂点なら，$v_{t,i}$ と $v_{t',j}$ の間の辺の重みはコスト関数の定義で用いられた α とする．さらに，それ以外の辺の重みは 0 とする．従って，グラフ系列 $\{G^{(1)},\cdots,G^{(T)}\}$ を G' に変換することで，グラフ系列のクラスタリング問題は G' のクラスタリング問題として解くことができる．

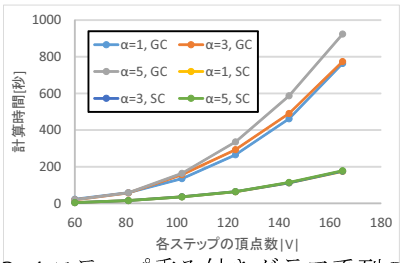

図 2: 4 ステップ重み付きグラフ系列の例

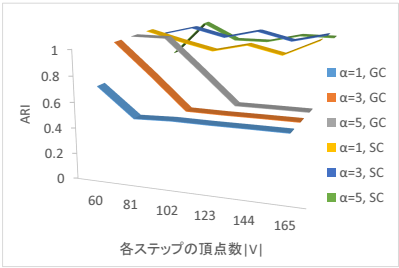

図 3: 4 ステップ重み付きグラフ系列の例

また，スペクトルクラスタリング [4] には外れ値の事例からなるサイズの小さいクラスタを抑制する効果のあることが知られている．提案手法では G' をスペクトルクラスタリングを用いて分割する．

4 評価実験

前節までに述べた提案手法を実装し，文献 [1] と同様に人工データを作成し，評価実験を行った．図 2 と図 3 は，それぞれ各ステップの頂点数 $|V|$ を変化させたときの計算時間と ARI (Adjusted Rand Index) である．また，GC と SC はそれぞれ G' をグラフカットとスペクトルクラスタリングで分割したことを表している．頂点数 $|V|$ が増加すると，G' の頂点数も増加するため計算時間が増加するが，SC を用いた場合の計算時間は GC を用いた場合と比べそれほど増加しない．なお，グラフ系列を G' に変換せず劣モジュラー最小化でグラフ系列をクラスタリングした場合は GC よりも非常に多くの計算時間を要する．また，$|V|$ が増加すると，数点の外れ値が発生するため最適解を求める GC の ARI は減少する．一方，SC はサイズの小さいクラスタを抑制する効果のあるため，$|V|$ が増加しても，SC の ARI は大きくは変化しない．以上より，グラフ系列のクラスタリングに対して，スペクトルクラスタリングを用いる効果を確認できた．

5 まとめ

本稿ではグラフ系列をスペクトルクラスタリングを用いて分割する手法を提案し，評価実験の結果を示した．

参考文献

[1] 岸本．劣モジュラ最適化に基づいたグラフ系列のクラスタリング．人工知能学会全国大会 1P2-lb-4in, 2011.

[2] 岸本．劣モジュラ最適化に基づいたグラフ系列のクラスタリング．大阪大学大学院工学研究科 修士論文, 2012.

[3] M. Queyranne. A combinatorial algorithm for minimizing symmetric submodular functions. *Proc. of SODA*, 98–101, 1995.

[4] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395-416, 2007.