

ハイパーグラフ上の相関クラスタリング に対する近似アルゴリズム

福永 拓郎^{1,a)}

概要：相関クラスタリングとは、オブジェクト間の相関関係を辺重み付きグラフによって表現し、グラフの頂点集合を分割する最適化問題を解くことによってオブジェクトのクラスタリングを行う手法のことである。本研究では、3つ以上のオブジェクト間の相関関係を扱うために、ハイパーグラフ上での相関クラスタリングを考える。これに伴って現れるハイパーグラフの分割問題に対して、線形計画法を利用した二つの近似アルゴリズムを提案する。一つ目のアルゴリズムは任意の重み付きハイパーグラフを扱うことができ、近似比 $O(r \log n)$ を達成する。ただし、 r は負の重みを持つ辺の最大サイズ、 n は頂点の数である。二つ目のアルゴリズムは、各辺の重みの絶対値が一定であり、各辺のサイズがちょうど k であるような k 部完全ハイパーグラフに対してのみ適用可能で、近似比 $O(k)$ を達成する。二つ目のアルゴリズムが対象となるようなハイパーグラフは、共クラスタリングの文脈で現れる。

Approximation algorithm for hypergraph correlation clustering

FUKUNAGA TAKURO^{1,a)}

Abstract: Correlation clustering is an approach for clustering a set of objects from given pairwise information. In this paper, we extend the correlation clustering to deal with higher-order relationships represented by hypergraphs. We give two approximation algorithms based on a linear programming relaxation of the problem. One achieves an $O(r \log n)$ -approximation, where n is the number of nodes and r is the maximum size of hyperedges with negative weights. This algorithm can be applied to hyperedges with arbitrary weights. The other achieves an $O(k)$ -approximation for complete k -uniform k -partite hypergraphs with weights of uniform absolute values. This type of hypergraphs arise from coclustering setting of correlation clustering.

1. はじめに

相関クラスタリングとは、オブジェクト間の相関関係からクラスタリングを行うための手法の一種である。この問題では、与えられた相関関係を無向グラフによって表現する。グラフの頂点はクラスタリングの対象となるオブジェクトに対応し、各辺は重みを持つ。正の重みを持つ辺は、その両端の頂点に対応するオブジェクトが同じクラスに属することを示唆し、負の重みを持つ辺はその両端点に対応するオブジェクトが異なるクラスに属することを意味する。ただし、ノイズの存在や観察の誤りなどのために、

辺の重みによって表現された二頂点間の関係は矛盾した情報を含むことを許す。相関クラスタリングでは、このように与えられた相関関係になるべく矛盾のないクラスタリングを、グラフの頂点集合分割問題を解くことによって求める。この問題は Bansal, Blum, Chawla [5] によって提案された。ほとんどのケースで NP 困難であることから様々な近似アルゴリズムがこれまで提案されており、機械学習や計算生物学で現れる応用に適用されるなどの成果も上がっている [6], [16], [21], [32]。

本研究の目的は、3つ以上のオブジェクト間の相関関係に関する情報も扱えるよう、相関クラスタリングの枠組みを拡張することである。3つ以上のオブジェクト間の関係からのクラスタリングは、2つのオブジェクト間の関係だけでは正確なクラスタリングを行うため必要な情報を得られ

¹ 国立情報学研究所, JST, ERATO, 河原林巨大グラフプロジェクト

^{a)} takuro@nii.ac.jp

ない場合に重要となる。例えば、ある空間に配置された点の集合を直線に分割するのはコンピュータビジョンにおいて重要なタスクであるが、複数の点が同一の直線に属しているかを判断するためには少なくとも3つ以上の点の位置関係を比較する必要がある、2つのデータ間の関係は全く役に立たない。この問題の従来手法では、全ての3点間の位置関係からクラスタリングを構築している [10], [28], [29]。

2つのオブジェクト間の相関関係がグラフで表現できるのに対して、3つ以上のオブジェクト間の関係はグラフを拡張したデータ構造であるハイパーグラフによって表現できる。グラフでは各辺は2つの頂点を結ぶのに対して、ハイパーグラフでは辺が3つ以上の頂点を結ぶことも許す。グラフに基づくクラスタリングアルゴリズムをハイパーグラフに拡張する試みはすでに多く存在する（詳しくは2章で述べる）。相関クラスタリングに関しても、画像分割への応用のために、Kim ら [26] によってハイパーグラフ上での相関クラスタリングがすでに研究されている。彼らはハイパーグラフ上の相関クラスタリングに対するアルゴリズムを提案し、これを利用することで画像分割の精度が向上することを実験を通して示した。ただし、彼らの提案アルゴリズムは近似精度に関する保証は与えられていない。本研究の目的の一つは、ハイパーグラフ上での相関クラスタリングに対して近似精度保証を持つアルゴリズムを提案することである。

本研究では一般のハイパーグラフ上での相関クラスタリングに加えて、双クラスタリング設定を拡張したときに現れるハイパーグラフにも着目する。双クラスタリングでは、例えば文章と単語、個人とグループ、映画と視聴者といったように、二種類のオブジェクトを分類するタスクのことである。ここで、分類のために利用可能な情報は異なる種類のオブジェクト間の関係であるとする。例えば、複数の文章と、それらの文書での単語の出現頻度から、文章と単語のクラスタリングを行うようなタスクが双クラスタリングに含まれる。また、遺伝子の発現データのクラスタリングにも有効であることが知られている。相関クラスタリングでは、双クラスタリングの設定はグラフが二部グラフである場合に対応し、多くの先行研究がある [2], [3], [13]。

3種類以上のオブジェクトのクラスタリングを2種類のオブジェクトのクラスタリングと区別するために、本稿では前者を共クラスタリング、後者を双クラスタリングと呼ぶことにする。共クラスタリングは多くの応用を持つ。例えば、オンラインショッピングの購買データから顧客、商品、ウェブサイト上で商品を検索した際のキーワードの3種類のオブジェクトを分類することを考える。もし顧客 i がキーワード j で検索した後に商品 k を購入した場合、 i, j, k は同じようなカテゴリーに属するだろうと考えられる。このような3者の関係を用いることで、より正確な分類を行うことができるのではないかと期待できる。

共クラスタリングは特に、特殊なハイパーグラフ上でのクラスタリングに対応する。オブジェクトが k 個のクラス $V_1, \dots, V_k$ に別れており、各 $j = 1, \dots, k$ について $i_j \in V_j$ であるような組み合わせ $(i_1, \dots, i_k)$ の相関関係からオブジェクト集合 $\bigcup_{i=1}^k V_i$ の分割を求める共クラスタリングを k 階共クラスタリングと呼ぶことにしよう。この場合、対応するハイパーグラフのそれぞれの辺は各 $j = 1, \dots, k$ について V_j の頂点をちょうど一つ含んでいる。そのようなハイパーグラフのことを本稿では k -ハイパーグラフと呼ぶ。

1.1 貢献

本研究では、ハイパーグラフ上での相関クラスタリングに対するピボッティングアルゴリズムに対して、近似精度保証を与える。ピボッティングアルゴリズムとは、以下の手順を繰り返してクラスタリングを計算するアルゴリズムのことである。

- (i) ピボット頂点 v を選ぶ。
- (ii) v を含むクラスターを、ある距離空間の下で v に近い頂点の集合と定義する。
- (iii) グラフ（またはハイパーグラフ）から v を含むクラスターに分類された頂点を取り除く。

グラフ上での相関クラスタリングに対して知られている近似アルゴリズムのほとんどはこの種類のアルゴリズムである。ステップ (i) でのピボット頂点 v の選択や、ステップ (ii) での v を含むクラスターの正確な定義についてはアルゴリズムごとに様々な設定が考えられている。

本研究では、問題の線形計画緩和を利用したピボッティングアルゴリズムを考える。我々の利用する線形計画緩和では、与えられたハイパーグラフの頂点上の距離空間を決める変数を含む。線形計画緩和を解くことにより、正の重みを持つ辺に含まれている2頂点はより近くに、負の重みを持つ辺に含まれる2頂点はより遠くに配置されるよう、距離空間が最適化される。このように得られた距離空間に基づき、ピボッティングアルゴリズムのステップ (i) と (ii) が実行される。

我々は二つの近似精度保証を与える。一つ目は、一般のハイパーグラフに適用できる近似精度保証である。 n を頂点の数、 r を負の重みを持つ辺の最大サイズとすると、アルゴリズムが $O(r \log n)$ 近似解を必ず出力することを保証する。つまり、アルゴリズムが出力する解の目的関数値は、必ず最適値の $O(r \log n)$ 倍以下となる。二つ目の保証は、与えられたハイパーグラフが完全 k -ハイパーグラフで、辺の重みの絶対値が一定であるような場合にのみ適用でき、近似比は $O(k)$ である。この近似比は k が定数であれば定数であり、二部グラフ上での相関クラスタリングに対して知られている定数近似アルゴリズム [2], [3], [13] を一般化しているとみなすことができる。これらの二つの保証を達成するピボッティングアルゴリズムはほとんど同じように

定義されるが、ステップ (i) と (ii) の細部においてやや異なる定義となっている。

ハイパーグラフ上での相関クラスタリングに対しては、Kim ら [26] も線形計画緩和を用いたピボッティングアルゴリズムを提案している。彼らの用いた線形計画緩和は我々のものと似ているが少し異なっている。また、彼らはアルゴリズムの近似精度保証は与えていない。実際、彼らのアルゴリズムはいくらでも悪い精度の解を出力することがありうることは容易に確かめることができる。彼らのアルゴリズムと比べて我々の提案アルゴリズムは、より洗練された丸め手法を用いている。

1.2 構成

本稿の構成は以下の通りである。2 章では関連研究について紹介する。3 章ではハイパーグラフ上での相関クラスタリングを導入する。4 章では $O(r \log n)$ 近似アルゴリズムを与え、5 章では重みの絶対値が一定の場合の完全 k -ハイパーグラフに対する $O(k)$ 近似アルゴリズムを与える。6 章では結論を与える。

2. 関連研究

2.1 相関クラスタリング

相関クラスタリングは Bansal, Blum, Chawla [5] によって提案された。これまでの研究では主に 2 種類の目的関数について研究されている。一つは合意最大化で、この設定ではクラスタリングの結果と一致する枝の重みの合計を最大化するのが目的である。もう一つは違反最小化の設定で、この設定ではクラスタリングの結果とは一致しない枝の重みの合計を最小化する。Charikar, Guruswami, Wirth [11] は、合意最大化の設定に対して 0.7664 近似アルゴリズムを与えた。相違最小化に対しては、Charikar らと Demaine ら [18] によって $O(\log n)$ 近似アルゴリズムが与えられた。Demaine らはまた、相違最小化の設定がマルチカット問題と同値であることを示すことにより、相違最小化の下での相関クラスタリングに対して定数近似を達成することがユニークゲーム困難であることを証明した [12]。

合意最小化の設定についてはいくつかの特殊ケースについても研究が行われている。例えば、Amit [3], Ailon ら [2], Chawla ら [13] はグラフが完全二部グラフで、全ての辺重みの絶対値が一定である場合を考えている。彼らはいずれもこの場合について定数近似アルゴリズムを与えており、その中でも最良の近似比は Chawla らによって与えられた 3 である。Chawla らはまた、グラフが完全グラフである場合も考え、全ての辺重みの絶対値が一定であれば 2.06 近似、辺重みが三角不等式を満たせば 1.5 近似を達成するアルゴリズムを与えた。

これらの先行研究は全てグラフ上での相関クラスタリングに関するものである。相関クラスタリングの問題設定は

様々な形で拡張されている [4], [9], [31], [36]。しかし、その中でハイパーグラフ上の相関クラスタリングを扱っている先行研究は、我々の知る限り Kim ら [26] によるもののみである。

ハイパーグラフのクラスタリングを扱うもっとも単純な方法は、ハイパーグラフをグラフに変形してグラフに対するクラスタリング手法を適用することである。もっとも一般的な変形方法としては、ハイパーグラフの各辺をグラフのスターやクリークに置き換えるものがある。これらはそれぞれ、スター拡大やクリーク拡大と呼ばれている [1], [44]。グラフの正規化カットを求めるアルゴリズムはグラフ上でのクラスタリングに頻繁に用いられるが、この手法をハイパーグラフに拡張する試みはいくつか知られている [8], [27], [37], [41]。Bulo と Pelillo [10] はハイパーグラフのクラスタリングゲームを提案し、そのゲームの解からクラスタリングを計算するアプローチを提案した。このアプローチの後続研究としては Liu, Latecki, Yan [29] や Leordeanu, Sminchisescu [28] によるものが知られている。

2.2 共クラスタリング

行列によって表現されたデータの双クラスタリングは 1970 年代から研究されている [22]。これまで多くのアルゴリズムが提案されているが、ここではその中でも代表的なものをいくつか参照するに留める [20], [33], [38], [39]。これらのアルゴリズムは多くの教師なし学習のタスクに適用され成功を収めている [14], [17], [43]。特に、遺伝子の発現データのクラスタリング [15], [24], [30] や文章の分類 [7], [19], [25] に対する応用は活発に研究が行われている。双クラスタリングと比較すると、より高い階数を持つ関係データの共クラスタリングはこれまであまり研究がなされてこなかった。Zhao と Zaki [40] はグラフを用いた共クラスタリングアルゴリズムを提案している。Hatano, Fukunaga, Kawarabayashi [23] はハイパーグラフのマルチカットのサンプリングに基づいた共クラスタリングアルゴリズムを提案している。その他の先行研究 [34], [35], [42] は、テンソル分解などの代数的な手法に基づいている。

3. 準備

3.1 問題設定

V を頂点集合とする。本稿では、 n で V の大きさ $|V|$ を表す。頂点集合 V 上のハイパーグラフでは、辺は V の部分集合として定義される。 E を辺集合とし、 $G = (V, E)$ で頂点集合 V 、辺集合 E からなるハイパーグラフを表す。辺 e の大きさを e のランクと呼び、ハイパーグラフ G のランクを G に含まれる辺のランクの最大値と定義する。ランク 2 のハイパーグラフは通常のグラフとなる。 V の部分集合 U と E の部分集合 H に対し、 H に含まれる辺のうち U と $V \setminus U$ の両方の頂点を含むものからなる集合を $\delta(U; H)$ で

表し、 $V \setminus U$ と交わらない H の辺の集合のことを $H[U]$ で表す。また、 $G[U]$ を U によって誘導される G の部分ハイパーグラフ、つまり頂点集合 U 、辺集合 $E[U]$ からなるハイパーグラフとする。

ハイパーグラフ上の相関クラスタリングでは、ハイパーグラフ $G = (V, E)$ が与えられる。1 章では相関クラスタリングの導入の際、グラフの各辺が正負両方の値をとりうる重みを持つとしたが、以降は G の各辺は正の重みとラベルを持つとする。辺のラベルは正か負のどちらかのラベルをとる。 E_+ と E_- でそれぞれ、正のラベル、負のラベルを持つ辺の集合を表すとする。 E の各辺は E_+ か E_- のどちらかに所属する。各辺の重みは正の実数として定義される。辺 e の重みを $w(e)$ と記す。また、 E_- に含まれる辺の最大ランクを r で記す。 $G = (V, E)$ のクラスタリング $\mathcal{C}$ は V の分割として定義される。つまり、 $\mathcal{C}$ は V の互いに素で非空な部分集合からなり、 V の各頂点は $\mathcal{C}$ に含まれる部分集合のいずれかちょうど一つに含まれる。相関クラスタリングでは、与えられたハイパーグラフのクラスタリングのうち、最適な目的関数値を達成するものを求める問題である。

2 章で触れたように、相関クラスタリングの目的関数としては合意最大化と違反最小化がこれまで考えられてきた。本稿では、違反最小化を扱う。得られたクラスタリングの目的関数値は、ハイパーグラフの各辺の重みやラベルによって表現される頂点間の相関関係の情報とどれだけ異なっているかによって決まる。正のラベルを持つ辺は、その辺に含まれる頂点は同じクラスに所属することを意味し、負のラベルを持つ辺はその辺に含まれる頂点は全ておなじクラスに所属する訳ではないことを意味する。辺の重みは、その情報の信頼度を表現する。グラフ上での相関クラスタリングでは、次のいずれかの条件を満たす時、クラスタリング $\mathcal{C}$ は辺 e に違反すると呼ばれる。

- 辺 e は正のラベルを持ち、その両端点はクラスタリング $\mathcal{C}$ において異なるクラスに分類される。
- 辺 e は負のラベルを持ち、その両端点はクラスタリング $\mathcal{C}$ において同じクラスに分類される。

目的関数値は、そのクラスタリングが違反した辺の重みの合計として定義される。

ハイパーグラフ上の相関クラスタリングでは、任意の $l \geq 1$ について、正のラベルを持つ辺に含まれる頂点が $l+1$ 以上のクラスに分類されているときや、負のラベルを持つ辺の頂点が高々 l 以下のクラスに分類されている時に、クラスタリングがその辺に違反すると定義できる。我々は、 $l = 1$ の場合の定義を採用する。つまり、正のラベルを持つ辺が一つ以上のクラスと交わる時、負のラベルを持つ辺に含まれる全ての頂点がが一つのクラスに含まれる時、クラスタリングがその辺に違反すると定義する。

1 章で触れたように、オブジェクトが k 個のクラスに分

かれている場合の共クラスタリングは、 k -ハイパーグラフに対応する。本稿では、 k -ハイパーグラフでの頂点集合の分割を $\{V_1, \dots, V_k\}$ で表す。与えられたハイパーグラフが完全 k -ハイパーグラフで、全ての辺の重みが一定の場合を議論する。つまり、辺集合は $V_1, \dots, V_k$ それぞれからちょうど 1 頂点ずつを含む辺からなり、全ての辺の重みは 1 であるとしてよい。

3.2 線形計画法に基づくピボットングアルゴリズム

本節では、ハイパーグラフ上の相関クラスタリングに対する提案アルゴリズムを導入する。提案アルゴリズムは、問題を整数計画問題としての定式化を緩和して得られる線形計画問題に基づいている。まず最初に、整数計画問題としての定式化を与える。この定式化では、0 か 1 の値をとる以下の変数を用いる。

- 各頂点对 $u, v \in V$ について定義される変数 x_{uv} 。頂点 u と v が同じクラスに分類される時 $x_{uv} = 0$ となり、異なるクラスに分類される時 $x_{uv} = 1$ となる。
- 各辺 $e \in E$ について定義される変数 x_e 。 e に含まれる全ての頂点が同じクラスに分類される時 $x_e = 0$ となり、2 つ以上のクラスに分類される時 $x_e = 1$ となる。

正のラベルをもつ辺 $e \in E_+$ が一つのクラスに含まれることは、 e に含まれる任意の 2 頂点は同じクラスに分類されることを意味する。これは以下のような条件によって表現できる。

$$x_{uv} \leq x_e, \forall e \in E_+, \forall \{u, v\} \subseteq e. \quad (1)$$

負のラベルを持つ辺 $e = \{v_1, \dots, v_r\} \in E_-$ が一つ以上のクラスと交わっているとき、頂点 v_1 の所属するクラスと $v_2, \dots, v_r$ のいずれかの頂点のクラスは異なっているはずである。このことより、以下の条件が成り立つ。

$$x_e \leq \sum_{i=1}^{r-1} x_{v_i v_{i+1}}, \forall e = \{v_1, \dots, v_r\} \in E_-. \quad (2)$$

ここで、 e に含まれる頂点の順序は任意の順序をとるとする。

2 頂点 u, v が同じクラスに分類され、 v と z も同じクラスに分類される時には、これらの 3 つの頂点 u, v, z は全て同じクラスに所属するはずである。このことは、 $x_{uv} = x_{vz} = 0$ である場合には $x_{uz} = 0$ が成立することを意味する。よって、これらの変数は以下の三角不等式を満たす。

$$x_{uz} \leq x_{uv} + x_{vz}, \forall u, v, z \in V. \quad (3)$$

我々の整数計画問題では、これらの制約のもとで目的関数の最適化を行う。違反最小化目的関数は

$$\sum_{e \in E_+} w(e)x_e + \sum_{e \in E_-} w(e)(1 - x_e)$$

と定義できる。

提案アルゴリズムでは、この整数計画問題での各変数の値域を $[0, 1]$ 上に緩和することによって得られる線形計画問題を用いる。つまり、線形計画問題は以下になる。

$$\begin{aligned} \text{最小化} \quad & \sum_{e \in E_+} w(e)x_e + \sum_{e \in E_-} w(e)(1 - x_e) \\ \text{制約} \quad & (1), (2), (3), \\ & x_{uv} \in [0, 1], \quad \forall u, v \in V, \\ & x_e \in [0, 1], \quad \forall e \in E. \end{aligned} \quad (4)$$

また、問題 (4) には現れないが、利便性のため全ての $v \in V$ に対して変数 $x_{vv} = 0$ を定義する。

ピボットリングアルゴリズムでは、まず最初に線形計画緩和問題 (4) の最適解 x を計算する。次に、1.1 章で述べた 3 つのステップを反復することで、 x からクラスタリングを構築する。 U を、ある反復の始まりでどのクラスタにも所属しない頂点の集合とする。この反復では、 U からピボット頂点 v を選ぶ。その後、 v を含むクラスタを、与えられる半径 $\xi \in [0, 1]$ を元に $B_{x,v,U}(\xi) := \{u \in U : x_{uv} < \xi\}$ と定義する。我々の提案アルゴリズムでは、 $B_{x,v,U}(\xi)$ と交わっている辺のうち違反したものの重みの合計と $B_{x,v,U}(\xi)$ と交わっている全ての辺の重みの合計の比を最小化するようにピボット頂点や半径を設定する。詳細は、4 章と 5 章に記述する。アルゴリズムの全体像はアルゴリズム 1 に記す。

アルゴリズム 1 線形計画緩和に基づくピボットリングアルゴリズム

入力: ハイパーグラフ $G = (V, E)$, E の分割 $\{E_+, E_-\}$, 各 $e \in E$ の非負重み $w(e)$

出力: クラスタリング $\mathcal{C}$

- 1: $\mathcal{C} \leftarrow \emptyset, U \leftarrow V$
- 2: (4) の最適解 x を計算
- 3: **while** $U \neq \emptyset$ **do**
- 4: ピボット頂点 $v \in U$ と半径 $\xi \in [0, 1]$ を計算 (詳細は 4 章と 5 章)
- 5: $\mathcal{C} \leftarrow \mathcal{C} \cup \{B_{x,v,U}(\xi)\}, U \leftarrow U \setminus B_{x,v,U}(\xi)$
- 6: $B_{x,v,U}(\xi)$ と交わる辺を E から除去
- 7: **end while**
- 8: $\mathcal{C}$ を出力

4. $O(r \log n)$ 近似アルゴリズム

本章では、任意の辺重みをもつ一般のハイパーグラフ上での相関クラスタリングを考える。まず、いくつかの記号を導入する。問題 (4) の最適解とその目的関数値を x と L で表す。辺 e と頂点 v について、 $\min_{u \in e} x_{vu}$ と $\max_{u \in e} x_{vu}$ を達成する頂点をそれぞれ $n(e, v)$ と $f(e, v)$ で表す。半径 $\xi \in [0, 1]$ に対し、 $F_{v,x,E}(\xi)$ と $C_{v,x,E}(\xi)$ を次のように定義する。

$$\begin{aligned} F_{v,x,E}(\xi) &:= \frac{L}{n} + \sum_{e \in E_+ \cap [B_{x,v,U}(\xi)]} w(e)x_e \\ &+ \sum_{e' \in \delta(B_{x,v,U}(\xi); E_+)} w(e')x_{e'} \frac{\xi - x_{vn(e',v)}}{x_{vf(e',v)} - x_{vn(e',v)}}, \\ C_{v,x,E}(\xi) &:= \sum_{e \in \delta(B_{x,v,U}(\xi); E_+)} w(e). \end{aligned}$$

ただし、 $\delta(B_{x,v,U}(\xi); E_+) = \emptyset$ である場合 $C_{v,x,E}(\xi)$ は $+\infty$ と定義することとする。もし辺 e' が $e \in \delta(B_{x,v,U}(\xi); E_+)$ を満たすならば、 $n(e', v) \leq \xi \leq f(e', v)$ が成立するので、 $F_{v,x,E}(\xi)$ の第 3 項は高々 $\sum_{e' \in \delta(B_{x,v,U}(\xi); E_+)} w(e')x_{e'}$ である。以下では、文脈から明らかな時には $B_{x,v,U}(\xi)$, $F_{v,x,E}(\xi)$, $C_{v,x,E}(\xi)$ の添え字を省略する。

提案アルゴリズムの詳細について説明する。このアルゴリズムでは、ピボット頂点 v は U に含まれる任意の頂点として定義される。ピボット頂点 v を含むクラスタ $B(\xi)$ を定義する半径 ξ は $[0, 1/(2r)]$ 上で $C(\xi)/F(\xi)$ を最小化するように設定する。そのような ξ を求める問題は一種の連続最適化問題であるが、以下の理由から $O(n)$ 通りの値の中から最もよいものを選べば十分である。 U に含まれる頂点を $u_1, \dots, u_{|U|}$ とし、一般性を失うことなく $x_{vu_1} \leq x_{vu_2} \leq \dots \leq x_{vu_{|U|}}$ が成り立つとする。また、 $i \in \{1, \dots, |U| - 1\}$ とする。任意の $\xi', \xi'' \in (x_{vu_i}, x_{vu_{i+1}}]$ に対して $B(\xi') = B(\xi'')$ が成立することから、 $\xi' \leq \xi''$ であれば $F(\xi') \leq F(\xi'')$ と $C(\xi') = C(\xi'')$ が成り立つ。これらの二つの関係は $C(\xi')/F(\xi') \geq C(\xi'')/F(\xi'')$ を意味している。よって、 $C(\xi)/F(\xi)$ を最小化する ξ は、 $\{0, 1/(2r)\} \cup \{x_{vu} : u \in U, x_{vu} \leq 1/(2r)\}$ に含まれることがわかる。この実数値集合のサイズは $O(|U|) = O(n)$ である。

我々のアルゴリズムの近似比は $C(\xi)/F(\xi)$ に依存する。正確には、もし任意の反復で $C(\xi)/F(\xi) \leq \alpha$ が成立するならば、アルゴリズムの近似比は $2 \max\{r, \alpha\}$ となる。以下の補題 1 は、常に $C(\xi)/F(\xi) \leq 2r \log(n+1)$ を満たす半径 $\xi \in [0, 1/(2r)]$ が存在することを示している。

補題 1. 任意の頂点 $v \in U$ に対し、 $C_{v,x,E}(\xi) \leq 2r \log(n+1) F_{v,x,E}(\xi)$ を満たす半径 $\xi \in [0, 1/(2r)]$ が存在する。

Proof. $\xi \in (0, 1/(2r))$ とする。ある頂点 $u \in V$ について $\xi = x_{vu}$ でなければ $F(\xi)$ は微分可能である。 $(0, 1/(2r))$ 上で $F(\xi)$ が微分可能でないときの ξ の値を $a_1, \dots, a_l$ とし、一般性を失うことなく $0 < a_1 < a_2 < \dots < a_l < 1/(2r)$ とする。また、 a_0 と a_{l+1} をそれぞれ $0, 1/(2r)$ と定義する。

$i \in \{0, \dots, l\}$ とする。制約 (1) から $x_e \geq x_{n(e,v)f(e,v)}$ が、制約 (3) から $x_{n(e,v)f(e,v)} \geq x_{vf(e,v)} - x_{vn(e,v)}$ が成り立つことから、 $x_e \geq x_{vf(e,v)} - x_{vn(e,v)}$ が成立する。よって、任意の $\xi \in (a_i, a_{i+1})$ について

$$\frac{d}{d\xi} F(\xi) = \sum_{e \in E_+ \cap \delta B(\xi)} w(e) \frac{x_e}{x_{vf(e,v)} - x_{vn(e,v)}} \geq C(\xi)$$

となる。

ϵ を正の実数とし、任意の $\xi \in [0, 1/(2r)]$ に対して $C(\xi) > \epsilon F(\xi)$ が成り立つと仮定しよう。この場合、 $\epsilon < r \log(n+1)$ となることを証明する。 $\frac{d}{d\xi} F(\xi) > C(\xi)$ と $C(\xi) > \epsilon F(\xi)$ から、全ての $\xi \in (a_i, a_{i+1})$ について $\frac{d}{d\xi} F(\xi) > \epsilon F(\xi)$ となる。よって、

$$\int_{F(a_i)}^{F(a_{i+1})} \frac{1}{F(\xi)} dF(\xi) > \int_{a_i}^{a_{i+1}} \epsilon d\xi = \epsilon(a_{i+1} - a_i).$$

が成り立つ。この不等式の左辺は

$$\int_{F(a_i)}^{F(a_{i+1})} \frac{1}{F(\xi)} dF(\xi) = \log F(a_{i+1}) - \log F(a_i).$$

である。これより、

$$\log F(a_{i+1}) - \log F(a_i) > \epsilon(a_{i+1} - a_i).$$

が導かれる。この不等式を全ての $i = 0, \dots, l$ について足し合わせることで、

$$\log F\left(\frac{1}{2r}\right) - \log F(0) > \frac{\epsilon}{2r}$$

が得られる。 $F(0) = L/n$ と $F(1/(2r)) \leq (1 + 1/n)L$ が成り立つことから、 $\epsilon < 2r \log(n+1)$ が得られる。□

定理 1. ハイパーグラフ上の相関クラスタリングに対する $4r \log n$ 近似アルゴリズムが存在する。

スペースの都合上、本稿では定理 1 の証明は省略する。

5. 一定重みの共クラスタリングに対する $O(k)$ 近似アルゴリズム

本章では共クラスタリングで一定の重みを持つ場合を考える。つまり、入力ハイパーグラフの頂点集合は互いに素な $V_1, \dots, V_k$ の集合和となっており、辺集合は $V_1 \times V_2 \times \dots \times V_k$ と定義され、各辺は重み 1 を持つ。問題の目的は、 $\{|e \in E_+ : e \in \delta(C)\}| + \{|e \in E_- : e \in E(C)\}|$ を最小化する $\bigcup_{i=1}^k V_i$ の分割 C を求めることである。本章を通して、 $k \geq 3$ を仮定する ($k=2$ の場合は Chawla ら [13] による 3 近似アルゴリズムが知られている)。

この場合に対する我々の提案アルゴリズムでは、 $U \cap V_1 \neq \emptyset$ である限りピボット頂点 v は $U \cap V_1$ から選ばれ、半径 ξ は $1/\sqrt{2(k-1)(2k-1)}$ に固定される。 $U \cap V_1 = \emptyset$ の時にはハイパーグラフには辺が残っていないので、 U に残された頂点の分類は問題の目的関数値には影響を与えない。よって、この場合にはアルゴリズムは反復を終了し、残りの頂点の任意の分割を解に加えたのち出力する。

ピボット頂点の選択についてより正確に説明するために、いくつか記号を導入する。以下では、記述の簡便さのため半径 ξ を $1/\theta$ と書き、 θ を固定されたパラメータとする。後に、 $\theta = \sqrt{2(k-1)(2k-1)}$ とすることで所望の近

似比を達成できることを示す。

1 を事象 Φ が真の時 $\mathbf{1}(\Phi) = 1$ 、偽の時 $\mathbf{1}(\Phi) = 0$ となるような関数とする。各反復で、頂点 $v \in U$ と残された辺 e について、二つのコスト $L_v(e)$ と $A_v(e)$ を以下のように定義する。

$$L_v(e) = \begin{cases} x_e \cdot \mathbf{1}(B(v, 1/\alpha) \cap e \neq \emptyset) & \text{if } e \in E_+, \\ (1 - x_e) \cdot \mathbf{1}(B(v, 1/\alpha) \cap e \neq \emptyset) & \text{if } e \in E_-, \end{cases}$$

$$A_v(e) = \begin{cases} \mathbf{1}(B(v, 1/\alpha) \cap e \neq \emptyset \neq e \setminus B(v, 1/\alpha)) & \text{if } e \in E_+, \\ \mathbf{1}(e \subseteq B(v, 1/\alpha)) & \text{if } e \in E_-. \end{cases}$$

もし頂点 v がこの反復でピボット頂点として選ばれると、線形計画緩和問題の最適値は少なくとも $\sum_{e \in E_+ \cup E_-} L_v(e)$ だけ減少し、保持している解の目的関数値は $\sum_{e \in E_+ \cup E_-} A_v(e)$ だけ増加する。アルゴリズムでは、各反復で $\sum_{e \in E_+ \cup E_-} A_v(e) / \sum_{e \in E_+ \cup E_-} L_v(e)$ を最小化するような頂点 $v \in V_1 \cap U$ をピボット頂点として選択する。

各反復のピボット頂点 v がある $\alpha \geq 1$ について $\sum_{e \in E_+ \cup E_-} A_v(e) / \sum_{e \in E_+ \cup E_-} L_v(e) \leq \alpha$ を満たすならば、アルゴリズムの近似比は高々 α である。以下では、 $\theta = \sqrt{2(k-1)(2k-1)}$ ならば、

$$\alpha = 2\sqrt{2(k-1)(2k-1)} + 4k - 3. \quad (5)$$

についてこの条件が成り立つことを示す。このためには、(5) を満たす α について

$$\sum_{v \in V_1 \cap U} \sum_{e \in E_+ \cup E_-} A_v(e) \leq \alpha \sum_{v \in V_1 \cap U} \sum_{e \in E_+ \cup E_-} L_v(e) \quad (6)$$

を示せば十分である。

以下では、簡単のため $U = V$ の仮定の下で (6) を証明する。 $U \subset V$ の場合は、以下の議論を U によって誘導される部分ハイパーグラフに適用すれば (6) が証明できる。

まず、次の補題で $\sum_{v \in V_1} \sum_{e \in E_-} A_v(e)$ の上界を与える。

$$\text{補題 2.} \quad \sum_{v \in V_1} \sum_{e \in E_-} A_v(e) \leq \frac{\theta}{\theta - 2k + 2} \sum_{v \in V_1} \sum_{e \in E_-} L_v(e).$$

次に $\sum_{v \in V_1} \sum_{e \in E_+} A_v(e)$ の上界を与える。このために、 $0 \leq \beta \leq 1/\theta$ を満たす実数 β を導入する。 $\theta \geq 2k - 2 \geq k$ より $1/\theta \leq (1 - 1/\theta)/(k - 1)$ が導かれるので、 β は $1 - 1/\theta - (k - 1)\beta \geq 0$ を満たす。以下の補題ではまず $\sum_{v \in V_1} \sum_{e \in E_+ : x(e) \geq \beta} A_v(e)$ を抑える。

$$\text{補題 3.} \quad \sum_{v \in V_1} \sum_{e \in E_+ : x(e) \geq \beta} A_v(e) \leq \frac{1}{\beta} \sum_{v \in V_1} \sum_{e \in E_+} L_v(e).$$

次に、 $\sum_{v \in V_1} \sum_{e \in E_+ : x(e) < \beta} A_v(e)$ の上界を考える。

$$\text{補題 4.} \quad \sum_{v \in V_1} \sum_{e \in E_+ : x(e) < \beta} A_v(e) \leq \frac{\theta}{1 - \theta\beta} \sum_{v \in V_1} \sum_{e \in E_+ \cup E_-} L_v(e).$$

補題 2, 3, 4 の証明はスペースの都合上本稿では省略する。これらの補題から以下が得られる。

$$\sum_{v \in V_1} \sum_{e \in E_+ \cup E_-} A_v(e) \leq \left(\max \left\{ \frac{\theta}{\theta - 2k + 2}, \frac{1}{\beta} \right\} + \frac{\theta}{1 - \theta\beta} \right) \sum_{v \in V_1} \sum_{e \in E_+ \cup E_-} L_v(e).$$

よって、 α が高々

$$\max \left\{ \frac{\theta}{\theta - 2k + 2}, \frac{1}{\beta} \right\} + \frac{\theta}{1 - \theta\beta} \quad (7)$$

であれば (6) が成り立つことが分かる。制約 $\theta \geq 2k - 2$, $0 \leq \beta \leq 1/\theta$ の下での (7) の最小値は $\theta = \sqrt{2(k-1)(2k-1)}$, $\beta = 1 - \sqrt{2(k-1)/(2k-1)}$ によって達成され、(5) の右辺と等しい。

定理 2. 辺重みが一定ならば、完全 k -ハイパーグラフ上での相関クラスタリングに対して近似比 $4k - 3 + 2\sqrt{2(k-1)(2k-1)}$ を達成するアルゴリズムが存在する。

6. 結論

本研究ではハイパーグラフ上での相関クラスタリングを考え、線形計画緩和に基づくピボットリングアルゴリズムについて二つの近似精度保証を与えた。一つは任意の辺重みを持つ一般のハイパーグラフに対して $O(r \log n)$ 近似を保証し、もう一つは共クラスタリングの設定において現れるハイパーグラフが一定の辺重みを持つ場合に対して $O(k)$ 近似を保証する。これらの二つの結果を比較すると、任意の辺重みを扱うことができる前者の結果の方が有用ではないかと考えられる。前者の設定では、辺重みのスケールを調整することにより得られるクラスタの数を調整することができるからである。また、後者の設定ではデータの欠損などを直接扱うことができないという欠点もある。ただし、後者は先行研究 [2], [3], [13] で議論された二部グラフ上での相関クラスタリングをより一般化した設定を扱っているという点では興味深い。

参考文献

- [1] Agarwal, S., Branson, K. and Belongie, S. J.: Higher order learning with graphs, *Machine Learning, Proceedings of the Twenty-Third International Conference*, pp. 17–24 (2006).
- [2] Ailon, N., Avigdor-Elgrabli, N., Liberty, E. and van Zuylen, A.: Improved approximation algorithms for bipartite correlation clustering, *SIAM Journal on Computing*, Vol. 41, No. 5, pp. 1110–1121 (2012).
- [3] Amit, N.: The Bicluster Graph Editing Problem, PhD Thesis, Tel Aviv University (2004).
- [4] Anava, Y., Avigdor-Elgrabli, N. and Gamzu, I.: Improved theoretical and practical guarantees for chromatic correlation clustering, *Proceedings of the 24th International Conference on World Wide Web*, pp. 55–65 (2015).
- [5] Bansal, N., Blum, A. and Chawla, S.: Correlation clustering, *Machine Learning*, Vol. 56, No. 1-3, pp. 89–113 (2004).
- [6] Ben-Dor, A., Shamir, R. and Yakhini, Z.: Clustering gene expression patterns, *Journal of Computational Biology*, Vol. 6, No. 3/4, pp. 281–297 (1999).
- [7] Bisson, G. and Hussain, S. F.: Chi-Sim: A new similarity measure for the co-clustering task, *Proceedings of the Seventh International Conference on Machine Learning and Applications*, pp. 211–217 (2008).
- [8] Bolla, M.: Spectra, Euclidean representations and clusterings of hypergraphs, *Discrete Mathematics*, Vol. 117, No. 1-3, pp. 19–39 (1993).
- [9] Bonchi, F., Gionis, A., Gullo, F. and Ukkonen, A.: Chromatic correlation clustering, *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1321–1329 (2012).
- [10] Bulò, S. R. and Pelillo, M.: A game-theoretic approach to hypergraph clustering, *Advances in Neural Information Processing Systems 22: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems 2009*, pp. 1571–1579 (2009).
- [11] Charikar, M., Guruswami, V. and Wirth, A.: Clustering with qualitative information, *Journal of Computer and System Sciences*, Vol. 71, No. 3, pp. 360–383 (2005).
- [12] Chawla, S., Krauthgamer, R., Kumar, R., Rabani, Y. and Sivakumar, D.: On the hardness of approximating multicut and sparsest-cut, *Computational Complexity*, Vol. 15, No. 2, pp. 94–114 (2006).
- [13] Chawla, S., Makarychev, K., Schramm, T. and Yaroslavtsev, G.: Near Optimal LP rounding algorithm for correlation clustering on complete and complete k -partite graphs, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 219–228 (2015).
- [14] Chen, X., Ritter, A., Gupta, A. and Mitchell, T. M.: Sense discovery via co-clustering on images and text, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5298–5306 (2015).
- [15] Cheng, Y. and Church, G. M.: Biclustering of expression data, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103 (2000).
- [16] Cohen, W. W. and Richman, J.: Learning to match and cluster large high-dimensional data sets for data integration, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 475–480 (2002).
- [17] Dao, P., Colak, R., Salari, R., Moser, F., Davicioni, E., Schönhuth, A. and Ester, M.: Inferring cancer sub-network markers using density-constrained biclustering, *Bioinformatics*, Vol. 26, No. 18 (2010).
- [18] Demaine, E. D., Emanuel, D., Fiat, A. and Immorlica, N.: Correlation clustering in general weighted graphs, *Theoretical Computer Science*, Vol. 361, No. 2-3, pp. 172–187 (2006).
- [19] Dhillon, I. S.: Co-clustering documents and words using bipartite spectral graph partitioning, *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 269–274 (2001).
- [20] Dhillon, I. S., Mallela, S. and Modha, D. S.: Information-theoretic co-clustering, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 89–98 (2003).
- [21] Filkov, V. and Skiena, S.: Integrating microarray data by consensus clustering, *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 418–425 (2003).

- [22] Hartigan, J. A.: Direct clustering of a data matrix, *Journal of the American Statistical Association*, Vol. 67, No. 337, pp. 123–129 (1972).
- [23] Hatano, D., Fukunaga, T. and Kawarabayashi, K.: Scalable algorithm for higher-order co-clustering via random sampling, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, to appear (2017).
- [24] Hussain, S. F.: Bi-clustering gene expression data using co-similarity, *Proceedings of the 7th International Conference on Advanced Data Mining and Applications*, pp. 190–200 (2011).
- [25] Hussain, S. F., Bisson, G. and Grimal, C.: An improved co-similarity measure for document clustering, *Proceedings of the Ninth International Conference on Machine Learning and Applications*, pp. 190–197 (2010).
- [26] Kim, S., Nowozin, S., Kohli, P. and Yoo, C. D.: Higher-order correlation clustering for image segmentation, *Advances in Neural Information Processing Systems 24: Proceedings of the 25th Annual Conference on Neural Information Processing Systems 2011*, pp. 1530–1538 (2011).
- [27] Leighton, T., Makedon, F. and Tragoudas, S.: Hypergraph partitioning algorithms, Technical Report PCS-TR94-233, Dartmouth College, Computer Science, Hanover, NH (1994).
- [28] Leordeanu, M. and Sminchisescu, C.: Efficient hypergraph clustering, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pp. 676–684 (2012).
- [29] Liu, H., Latecki, L. J. and Yan, S.: Robust clustering as ensembles of affinity relations, *Advances in Neural Information Processing Systems 23: Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pp. 1414–1422 (2010).
- [30] Madeira, S. C., Teixeira, M. C., Sá-Correia, I. and Oliveira, A. L.: Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, No. 1, pp. 153–165 (2010).
- [31] Mathieu, C., Sankur, O. and Schudy, W.: Online correlation clustering, *27th International Symposium on Theoretical Aspects of Computer Science*, pp. 573–584 (2010).
- [32] McCallum, A. and Wellner, B.: Toward conditional models of identity uncertainty with application to proper noun coreference, *Proceedings of IJCAI-03 Workshop on Information Integration on the Web, IIWeb-03*, pp. 79–84 (2003).
- [33] O'Connor, L. and Feizi, S.: Biclustering using message passing, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pp. 3617–3625 (2014).
- [34] Papalexakis, E. E., Sidiropoulos, N. D. and Bro, R.: From K-means to higher-way co-clustering: multilinear decomposition with sparse latent factors, *IEEE Transactions on Signal Processing*, Vol. 61, No. 2, pp. 493–506 (2013).
- [35] Peng, W. and Li, T.: Temporal relation co-clustering on directional social network and author-topic evolution, *Knowledge Information Systems*, Vol. 26, No. 3, pp. 467–486 (2011).
- [36] Rebagliati, N., Bulò, S. R. and Pelillo, M.: Correlation clustering with stochastic labellings, *Proceedings of the Second International Workshop on Similarity-Based Pattern Recognition*, pp. 120–133 (2013).
- [37] Rodríguez, J.: On the Laplacian spectrum and walk-regular hypergraphs, *Linear and Multilinear Algebra*, Vol. 51, No. 3, pp. 285–297 (2003).
- [38] Shan, H. and Banerjee, A.: Bayesian co-clustering, *Proceedings of the 8th IEEE International Conference on Data Mining*, pp. 530–539 (2008).
- [39] Zha, H., He, X., Ding, C. H. Q., Gu, M. and Simon, H. D.: Bipartite graph partitioning and data clustering, *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management*, pp. 25–32 (2001).
- [40] Zhao, L. and Zaki, M. J.: TriCluster: An effective algorithm for mining coherent clusters in 3D microarray data, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 694–705 (2005).
- [41] Zhou, D., Huang, J. and Schölkopf, B.: Learning with hypergraphs: clustering, classification, and embedding, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pp. 1601–1608 (2006).
- [42] Zhou, Q., Xu, G. and Zong, Y.: Web co-clustering of usage network using tensor decomposition, *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology — Workshops*, pp. 311–314 (2009).
- [43] Zhu, Y., Yang, H. and He, J.: Co-clustering based dual prediction for cargo pricing optimization, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1583–1592 (2015).
- [44] Zien, J. Y., Schlag, M. D. F. and Chan, P. K.: Multilevel spectral hypergraph partitioning with arbitrary vertex sizes, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 18, No. 9, pp. 1389–1399 (1999).