日本オペレーションズ・リサーチ学会
2023 年春季研究発表会

階層クラスタリングに対する許容的目的関数の特徴付けと関連する最適化問題に対する近似アルゴリズム

01013123　静岡大学　*安藤和敏　ANDO Kazutoshi
静岡大学　筑波竜希　TSUKUBA Ryuki

1. はじめに

ここで考えるクラスタリングとは, データ集合とデータの対ごとの類似度が与えられたときに, データ集合を類似するデータから成る部分集合 (クラスター) へと分割する手続きである．クラスタリングは，非階層クラスタリングと階層クラスタリングに大別される．非階層クラスタリングはデータ集合の予め定められた個数のクラスタへの分割を求めるのに対して，階層的クラスタリングは，データ集合のクラスタへの分割の階層構造を求める．非階層クラスタリングに対しては目的関数が定義されており，よく知られている k 平均法はこの目的関数を最小化する局所探索アルゴリズムである．階層クラスタリング・アルゴリズムには，単連結法や完全連結法のような凝集的方法と呼ばれる多くのアルゴリズムがよく知られているが，これらのアルゴリズムは何らかの目的関数値を最適化する階層クラスタリングを生み出すものではないため，これらの手法の優劣を客観的に比較することは困難であった．階層構造は葉集合がクラスター木と呼ばれる 2 分木 (デンドログラム) によって表現される．Dasgupta [4] は与えられた入力データに対する任意の階層クラスタリングを評価するために，クラスター木を引数とする目的関数を導入し，階層クラスタリングの問題をこの目的関数を最小化するクラスター木を求める組合せ最適化問題として定式化した．Dasgupta [4] はこの問題は NP 困難であることを示すと同時に，再帰的最疎カットアルゴリズムがこの問題に対する $\mathrm{O}(\phi \log n)$ -近似アルゴリズムであることを示した．ここで, $n = |X|$ であり，ϕ は最疎カット問題に対する近似アルゴリズムの近似精度である．再帰的最疎カットアルゴリズムの近似率は後に Charikar et al. [2] 及び Cohen-Addad et al. [3] によって $\mathrm{O}(\phi)$ に改善された．

Dasguputa は [4] の中でより一般的な目的関数のクラスを導入し，これらの目的関数を最小化する問題に対する近似アルゴリズムを与えている．Charikar et al. [2] は，再帰的最疎カットアルゴリズムがこの一般化された Dasguputa の目的関数の最小化問題に対する $\mathrm{O}(c_f(n)\phi)$-近似アルゴリズムであることを示した．ここで, c_f は目的関数に依存するパラメータである．一方で，Cohen-Addad et al. [3] は，許容的目的関数と呼ばれる階層クラスタリングに対する目的関数のクラスを定義してその特徴付けを与えた．許容的目的関数には Dasgupta [4]（一般化されていない）目的関数が含まれる．

本研究では，最初に 3 次以下の多項式を用いて定義される許容的目的関数に対する部分的な特徴付けを与える．次に, 再帰的最疎カットアルゴリズムはこのような許容的目的関数を最小化するクラスター木を求める問題に対する $\mathrm{O}(\phi)$-近似アルゴリズムであることを示す.

2. 階層クラスタリングに対する許容的目的関数

$(G = (X, E), w)$ を重み付き無向グラフとする．X は分析対象のデータ集合と解釈され，任意の $\{x, y\} \in E$ に対して $w(x, y)$ は x と y の類似度を表すと解釈される．$Y, Z \subseteq X$ に対して

$$w(Y, Z) = \sum_{\{y,z\} \in E, y \in Y, z \in Z} w(y, z) \tag{1}$$

と定義する．G の**クラスター木**は，葉集合が X であるような根付き二分木である．クラスター木 T の葉 $x, y \in X$ に対して，x, y の最小共通祖先を $\mathrm{lca}_T(x, y)$ と表す．また，T の任意の部分木 T' に対して, $L(T')$ によって T' の葉集合を表す．我々は G クラスター木 T の目的関数 $\Gamma(T)$ として，以下の (2) と (3) で定義されるものを考える．

$$\Gamma(T) = \textstyle\sum_v \gamma(v), \tag{2}$$

$$\gamma(v) = w(L(T_v^+), L(T_v^-)) \cdot g(|L(T_v^+)|, |L(T_v^-)|). \tag{3}$$

ここで, (2) 式の総和は T のすべての内部ノード v について取る．また (3) 式の T_v^+ と T_v^- は v の 2 つの子を根とする部分木であり，$g\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ は任意の関数である．Dasgupta の目的関数 [4] は $g(a, b) = a + b$ によって定義される．**階層クラスタリング最適化問題**とは，重み付きグラフ (G, w) が与えられたとき，$\Gamma(T)$ を最小化する G のクラスター木 T を見出す問題である．

$(G = (X, E), w)$ を重み付きグラフとする．G のクラスター木 T は，もし T の全ての内部ノード u, v に

対して u が v の親ならば $h(u) \leq h(v)$ であるような重み関数 h が存在して，すべての $x, y \in X$ に対して $w(x,y) = h(\mathrm{lca}_T(x,y))$ が成り立つならば，G の**生成木**と呼ばれる．

(2) 式と (3) 式の形式の目的関数 Γ は，条件「生成木が存在する任意の重み付きグラフ $(G=(X,E),w)$ に対して，クラスター木 T が Γ を最小化するための必要十分条件は T が G の生成木であることである」を満たすときに**許容的**と呼ばれる [3].

命題 2.1 (Cohen-Addad et al. [3]) Γ が許容的目的関数であるための必要十分条件は Γ が以下の性質を満たすことである．

(1) G がすべての辺の重みが 1 であるような完全グラフのとき，G のすべてのクラスター木 T に対する目的関数値 $\Gamma(T)$ は等しい．

(2) 各 $a, b \in \mathbb{N}$ に対して，$g(a,b) = g(a,b)$.

(3) 各 $a, b \in \mathbb{N}$ に対して，$g(a+1,b) > g(a,b)$.

Dasgupta の定義した目的関数 Γ は許容的である．

定理 2.2 $g(a,b)$ は 3 次以下の多項式とする．(2) と (3) によって定義される目的関数 Γ が許容的であるための十分条件は，$\lambda \geq 0, \mu \geq 0, 15\lambda + 9\mu + \nu > 0$ を満たす λ, μ, ν に対して

$$\begin{aligned} g(a,b) =& \lambda\left((a+b)^3 - (a+b)ab\right) \\ &+ \mu\left(2(a+b)^2 - ab\right) + \nu(a+b) \end{aligned} \tag{4}$$

が成り立つことである．

3. 近似アルゴリズム

$(G=(X,E),w)$ を重み付きグラフとする．G の**カット**とは G の点集合の 2 分割 $\{Y, X \setminus Y\}$ のことである．G のカット $\{Y, X \setminus Y\}$ の**密度** $d(Y, X\backslash Y)$ は

$$d(Y, X\backslash Y) = \frac{w(Y, X \setminus Y)}{|Y||X\backslash Y|} \tag{5}$$

によって定義される．**最疎カット問題**は最小の密度を持つ G のカットを求める問題であり，この問題は NP 困難である．**ϕ-最疎カット**とは G に対する最疎カット問題の ϕ-近似解のことである．例えば，Arora et al. [1] の近似アルゴリズムでは $\phi = \mathrm{O}(\sqrt{\log n})$ である．

定理 2.2 の条件を満たす任意の g 対して，

$$c_g(n) = \max_{2 \leq k \leq n} \frac{g(k-1,1)}{g\left(\lfloor \frac{k}{2} \rfloor /2, \lfloor \frac{k}{2} \rfloor /2\right) - g\left(\lfloor \frac{k}{4} \rfloor /2, \lfloor \frac{k}{4} \rfloor /2\right)} \tag{6}$$

と定義する．ここで，g の定義域は $\mathbb{R}_+ \times \mathbb{R}_+$ に拡大されている．

入力: 重み付きグラフ $(G=(X,E),w)$

1 $\{Y, X\backslash Y\}$ を G の ϕ-最疎カットとする;

2 誘導部分グラフ $G[Y]$ と $G[X\backslash Y]$ に対してアルゴリズムを再帰的に呼び出して，2 分木 T_Y と $T_{X\backslash Y}$ を得る;

3 **return** $T_Y, T_{X\backslash Y}$ 両方の根を子に持つノードを根とする木;

アルゴリズム 1: 再帰的最疎カットアルゴリズム．

定理 3.1 目的関数 Γ は定理 2.2 の条件を満たす g によって定義されると仮定する．アルゴリズム 1 は，入力として与えられた任意の重み付きグラフ $(G=(X,E),w)$ に対して，$\Gamma(T) \leq 8\phi c_g(n)\mathrm{OPT}$ であるようなクラスター木 T を出力する．ここで, OPT は (G,w) に対する階層クラスタリング最適化問題の最適値である．

$c_g(n) = \mathrm{O}(1)$ であるのでアルゴリズム 1 は階層クラスタリング最適化問題に対する $\mathrm{O}(\phi)$- 近似アルゴリズムである．

4. おわりに

本研究では，3 次以下の多項式を用いて定義される許容的目的関数に対する部分的な特徴付けを与え, 再帰的最疎カットアルゴリズムはこのような許容的目的関数を最小化するクラスター木を求める問題に対する $\mathrm{O}(\phi)$-近似アルゴリズムであることを示した．3 次の (あるいはより高次の) 多項式を用いて定義される許容的目的関数に対する必要十分条件を与えることとそれに対する近似アルゴリズムの導出は今後の課題である．

謝辞

本研究は JSPS 科研費 22K11921 の助成を受けたものである．

参考文献

[1] S. Arora et al.: Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM* **56** (2009) Article 5.

[2] M. Charikar et al.: Approximate hierarchical clustering via sparsest cut and spreading metrics. *SODA'17* (2017), pp. 841–854.

[3] V. Cohen-Addad et al.: Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM* **66** (2019) Article 26.

[4] S. Dasgupta: A cost function for similarity-based hierarchical clustering. *STOC '16* (2016), pp. 118-127.