2F3-05

疑似クリーク探索によるクラスター抽出

Extracting Clusters by Pseudo-Clique Search

大久保 好章　　原口 誠
Yoshiaki OKUBO　Makoto HARAGUCHI

北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Division of Computer Science, Graduate School of Information Science and Technology, Hokkaido University

In this paper, we are concerned with a problem of extracting clusters by finding *pseudo-cliques* in a given graph G. A pseudo-clique is defined as the union of several maximal cliques in G with a required degree of overlap. Such a degree is determined by a user-defined parameter τ. We present a depth-first algorithm for finding maximal pseudo-cliques whose sizes are in the top N. Based on some simple theoretical properties, effective pruning rules can be applied during our search. Our preliminary experimental result shows that some advantage of considering Top-N pseudo-cliques for cluster extraction. It is also verified that the prunings are invoked very frequently in the search.

1 はじめに

様々な応用領域における重要なタスクが，無向グラフにおける**最大クリーク**，あるいは，**極大クリーク**の抽出問題として定式化できることが知られている．著者らはこれまで，固体間の類似関係をグラフ表現し，そこからサイズが上位 N の極大クリーク，すなわち，Top-N 極大クリークを抽出することで，様々な**クラスター**を見つける枠組について考察してきた [3]．それにより，興味あるクラスターの抽出が可能であることが確認できた一方で，構成頂点がわずかに異なる極大クラスターが多数抽出され，それらが Top-N のほとんどを埋め尽くしてしまう現象も度々観測された．これは，本質的に異なるクラスターの数が N に対して極僅かであることを意味し，様々なクラスターを見つける立場からは望ましくない．この様な場合，重複の度合が大きな極大クラスター同士は敢えて区別せず，ひとつのクラスターを形成していると考えることで，本質的に異なるクラスターをより多く見つけることが可能となろう．こうした背景のもと，本研究では，重複の度合が閾値以上の極大クリーク族をひとつの**疑似クリーク**と見做し，サイズが上位 N の極大疑似クリークを抽出する問題，すなわち，*Top-N* **極大疑似クリーク問題**を定義し，その計算アルゴリズムを与えることで，疑似クリークとしてのクラスタ抽出を試みる．特にここでは，極大クリーク族の重複部分を，その疑似クリークの**核**と定め，それらが統合された根拠を明確にする．

2 準備

V を頂点集合，$E \subseteq V \times V$ を辺の集合とする無向グラフ G を $G = (V, E)$ と表す．グラフ G において，頂点 $v \in V$ と隣接する頂点の集合を $N_G(v)$ で表し，その要素 (頂点) 数，すなわち，$|N_G(v)|$ を G における v の**次数**と言う．これを $degree_G(v)$ で参照する場合もある．なお，文脈上明らかな場合は，単に $N(v)$ や $degree(v)$ と略記する．

グラフ G の任意の (異なる) 頂点間に辺が存在する時，G を**完全グラフ**と呼ぶ．

グラフ $G = (V, E)$ おいて，V の部分集合を V' とする．$G' = (V', E \cap V' \times V')$ で定義されるグラフを，G の**部分グラフ**と呼び，$G(V')$ と表記する．特に，G' が完全グラフである時，それは**クリーク**と呼ばれ，単に，その構成頂点集合 V' で表すものとする．また，その**サイズ**を $|V'|$ で定める．G のクリーク Q と Q' が，$Q \subset Q'$ の関係にある時，Q' を Q の**拡張** (*extension*) と呼ぶ．G のクリークのうち，包含関係のもとで極大なものを，**極大クリーク**と呼ぶ．特に，サイズが最大である極大クリークは**最大クリーク**と呼ばれる．一般に，最大クリークは一意に決まらないことに注意する．

連絡先: 大久保 好章・原口 誠：北海道大学大学院情報科学研究科コンピュータサイエンス専攻
〒 060-0814 札幌市北区北 14 条西 9 丁目
E-mail: { yoshiaki, mh }@ist.hokudai.ac.jp

3 疑似クリーク

本節では，核を考慮した疑似クリークの概念を導入し，その抽出問題を定義する．

まず，所与の無向グラフ G における**疑似クリーク**を以下の通り定義する．

定義 3.1 (疑似クリーク)
$\mathcal{C} = \{C_1, \ldots, C_m\}$ を G の極大クリーク族とする．頂点集合

$$pseudo(\mathcal{C}) = \bigcup_{C_i \in \mathcal{C}} C_i$$

を，**疑似クリーク** (*psuedo-clique*) と呼び，その**重複度** $overlap(\mathcal{C})$ を

$$overlap(\mathcal{C}) = \min_{C_i \in \mathcal{C}} \left\{ \frac{\left| \bigcap_{C_j \in \mathcal{C}} C_j \right|}{|C_i|} \right\}$$

と定める．特に，各 C_i の積，すなわち，$\bigcap_{C_i \in \mathcal{C}} C_i$ を $pseudo(\mathcal{C})$ の**核** (*core*) と呼ぶ．$pseudo(\mathcal{C})$ の**サイズ**は，$|pseudo(\mathcal{C})|$ とする． ■

定義より，極大クリーク族の重複度 α は，$0 \leq \alpha \leq 1$ の値をとり，値が大きくなるに従い重複の度合が増す．

重複が少ない極大クリーク同士を統合して疑似クリークと見做しても，まとまり，すなわち，クラスターとしての説得力には欠けるであろう．その意味で，疑似クリークの重複度は，それを構成する極大クリークをひとつに見做すに至った根拠を表すと考えられる．よってここでは，重複度の閾値を設け，閾値以上の重複度を有する疑似クリークのみを抽出の対象とする．

特に，サイズが上位 N の極大な疑似クリークを求めることで，根拠が明白な様々なクラスターを抽出することを考える．

定義 3.2 (Top-N τ-極大疑似クリーク問題)
G を無向グラフ，τ を重複度閾値とする．重複度が τ 以上の G の極大疑似クリーク (これを τ-疑似クリークと呼ぶ) のうち，サイズが上位 N であるものを求める問題を *Top-N τ-極大疑似クリーク問題*と呼ぶ． ■

4 Top-N 疑似クリーク探索アルゴリズム

本節では，無向グラフ $G = (V, E)$ と重複度閾値 τ について，G の Top-N τ-極大疑似クリークを求める計算手続きを与える．その詳細を述べる前に，疑似クリークの基本的性質を示す．

まず，クリークの拡張候補頂点を定義する．

定義 4.1 (クリークの拡張候補頂点)
Q を G のクリークとする．Q の任意の頂点に隣接する $v \in V$ を，Q の*拡張候補頂点*と呼ぶ．Q のすべての拡張候補頂点の集合を $cand(Q)$ と表記する． ■

定義より，以下が成立することが容易にわかる．

観察 4.1
Q と Q' を $Q \subseteq Q'$ なる G のクリークとする．この時，$cand(Q) \supseteq cand(Q')$ かつ $|Q| + |cand(Q)| \geq |Q'| + |cand(Q')|$ である． ■

G のクリーク Q について，$Q \subseteq C_{max}$ なる G の任意の極大クリーク C_{max} は，$cand(Q)$ 中のいくつかの頂点で Q を拡張することで得られる．ここで，疑似クリークの核はクリークであることに注意すると，Q を核とする疑似クリークのサイズは高々 $|Q|+|cand(Q)|$ であることがわかる．このことから，以下の性質が明らかとなる．

観察 4.2
Q を G のクリークとする．今，Top-N の疑似クリークが暫定的に見つかっているとし，その最小サイズを k と仮定する．$|Q| + |cand(Q)| < k$ である時，Q の任意の拡張 Q' を核とする疑似クリークは Top-N に成り得ない． ■

τ を重複度閾値とする．今，τ-疑似クリーク $\tilde{C}$ の核を Q とすると，$\tilde{C}$ は，$Q \subseteq C$ かつ $|Q|/|C| \geq \tau$ なる任意の極大クリーク C の和集合として得られる．この様な C について，$Q \cup D = C$ となる，部分グラフ $G(cand(Q))$ の極大クリーク D が存在することに注意しよう．つまり，疑似クリーク $\tilde{C}$ を得るためには，$|Q|/(|Q| + |D|) \geq \tau$ なる $G(cand(Q))$ の任意の極大クリーク D を列挙すれば十分である．

極大クリークの列挙 [2] は，一般に計算コストの高いタスクであるが，ここでは以下の理由により，計算負荷が抑えられるものと期待できる．

- 疑似クリークの核が小さい場合，それを構成可能な極大クリークのサイズも，重複度閾値に制約されて必然的に小さくなる．サイズの小さな極大クリークを探索する際は，深さを限定した探索が行えるため，その計算負荷はそれほど高くないと期待できる．
- 一方，疑似クリークの核 Q が大きくなるにつれて，$cand(Q)$ のサイズは単調に減少し，それに伴い部分グラフ $G(cand(Q)$ の頂点数も少なくなる．小さなグラフから極大クリークを列挙する負荷はそれほど大きくなく，現実的な計算コストを期待できる．

特に，前者は次の性質に支持される．

観察 4.3
G のクリーク Q について，Q を核とする τ-疑似クリーク $\tilde{C}$ の抽出を考える．部分グラフ $G(cand(Q)$ のクリーク D について，$|D| > (\frac{1}{\tau} - 1) \cdot |Q|$ であるならば，$\tilde{C}$ の抽出過程において D の任意の拡張を考慮する必要はない． ■

核を Q とする疑似クリーク $\tilde{C}$ の抽出にあたっては，一般に，$G(cand(Q))$ の 極大クリークを計算する必要があることを述べたが，次の場合は，その計算をせずに $\tilde{C}$ を同定することが可能である．

観察 4.4
G のクリークを Q，重複度閾値を τ とする．以下が成り立つ時，$Q \cup cand(Q)$ は Q を核とする τ-極大疑似クリークとなる．

- $(\frac{1}{\tau} - 1) \cdot |Q| \geq k$，ここで k は $G(cand(Q))$ における最大クリークの上限値である．
- 任意の $v \in cand(Q)$ について，$G(cand(Q))$ における v の次数は，$|cand(Q)| - 1$ より小さい． ■

前者により $Q \cup cand(Q)$ が τ-極大疑似クリークとなることが保証されるが，一般に，その核は Q の拡張となる．核が Q そのものであることは，後者により保証される．

最大クリークの上限値は，多くの最大クリーク抽出アルゴリズムにおいて利用されており，例えば [1] では，逐次近似彩色に基づく上限値が採用されている．

以上の議論から，無向グラフ G における Top-N τ-極大疑似クリークを，観察 4.2， 4.3， 4.4 に基づく枝刈り規則を利用した深さ優先探索によって抽出するものとする．アルゴリズムを図 1 にまとめる．

5 予備実験

Top-N 極大疑似クリークを抽出することで，様々なバリエーションのクラスターが獲得可能となることを示すために予備実験を行った．

疑似クリークの抽出を試みるグラフは，2,340 の頂点と 9,391 の辺で構成されている*1．いくつかの重複度閾値のもとで，Top-20 の疑似クリーク抽出を試みた．なお，厳密な極大クリークを Top-20 抽出した場合，それらは大部分の頂点を共有する 2 つの極大クリーク族に大別できた．すなわち，上位 20 の極大クリークを抽出したが，本質的に異なるものは 2 種類であった．

これに対して，重複度閾値 $\tau = 0.8$ のもとで Top-20 の疑似クリークを抽出した場合，大別して 5 種類のクラスターを得ることができた．それらの中には，少数の大きな極大クリークから構成される疑似クリークや，対照的に，比較的小さな極大クリークが多数集まった疑似クリークがある．ここで，厳密な Top-N 極大クリーク探索によって，後者の様な，重複の度合が大きく，かつ，サイズが小さな極大クリークが多数存在

*1 これは，ある生物の遺伝子発現時系列データを，その発現パターンの類似性に基づいてグラフ表現したものである．

```
procedure main() :
    V ← the set of vertices in a graph ;
    E ← the set of edges in the graph ;
    N ← an integer for Top-N ;
    τ ← a threshold for overlap degree ;
    PC ← φ;
    size_num ← 0 ;
    min_size ← 0 ;
    FindPseudoCliques(φ, V) ;
    return PC ;
```

```
procedure FindPseudoCliques(Q, R) :
    if size_num = N and |Q| + |R| < min_size then
        return ;    /* 観察 4.2 */
    endif
    for each v ∈ R in predetermined order
        begin
            MC ← φ ;
            α ← (1/τ − 1) · (|Q| + 1) ;
            k ← an upper bound of the max clique in G(R ∩ N(v)) ;
            if k ≤ α then
                if ∀w ∈ R ∩ N(v),
                  degree_{G(R∩N(v))}(w) < |R ∩ N(v)| − 1 then
                    MC ← {R ∩ N(v), φ} ;    /* 観察 4.4 */
                else
                    FindMaxCliques(φ, R ∩ N(v)) ;
                endif
            else
                FindMaxCliques(φ, R ∩ N(v)) ;
            endif
            if ⋂_{C_i∈MC} C_i = φ then
                if size_num < N or
                  |⋃_{C_i∈MC} C_i ∪ Q ∪ {v}| ≥ min_size then
                    PC ← PC ∪ {⋃_{C_i∈MC} C_i ∪ Q ∪ {v}} ;
                    size_num ← |{|PC| | PC ∈ PC}| ;
                    min_size ← min{|PC| | PC ∈ PC} ;
                endif
            endif
            FindPseudoCliques(Q ∪ {v}, R ∩ N(v)) ;
        end
```

```
procedure FindMaxCliques(Q, R) :
    if |Q| > α then
        return ; /* 観察 4.3 */
    endif
    if R = φ then
        MC ← MC ∪ {Q} ;
        return ;
    endif
    for each v ∈ R in predetermined order
        FindMaxCliques(Q ∪ {v}, R ∩ N(v)) ;
```

図 1: Top-N τ-疑似クリーク抽出アルゴリズム

することを認識するには，N を十分大きく設定する必要があることに注意しよう．その場合，膨大な極大クリークが出力され，我々はそれらの中からこうしたものを探し出すことを強いられる．しかし，疑似クリークを考えることで，そうした手間なしに，その存在を容易に認識することが可能となり，興味あるクラスターの獲得チャンスが広がるものと期待できよう．これは，疑似クリークを考えることによる極めて重要かつ有用な効果であることを強調しておく．

また，本予備実験により，探索過程において，本稿での枝刈り規則が十分な頻度で適用されることも確認できた．

6 おわりに

本稿では，クリーク探索によるクラスタ抽出の問題点を緩和すべく，疑似クリークの概念を導入し，深さ優先探索により Top-N 極大疑似クリークを抽出するアルゴリズムについて考察した．疑似クリークは，閾値以上の重複度を有する極大クリークの和集合で定義され，本質的には同じと見做せる極大クリークをひとつに統合したものに相当する．特にここでは，重複部分を疑似クリークの核と定めることで，構成する極大クリークを統合するに至った根拠を明確化した．予備実験により，疑似クリークを考えることで，これまで見つけることが困難であった様々なバリエーションのクラスター抽出が期待できることを確かめた．また，探索における枝刈りが有効に機能していることも確認した．なお，本アルゴリズムは，頂点が重みを有する重み付きグラフにおける**重み** *Top-N* **極大疑似クリーク探索**に対しても，容易に拡張可能である．

今後は，より大規模なグラフを扱える様，アルゴリズムのさらなる改良を行ないながら，実験を本格化する．現在，疑似クリーク探索による遺伝子発現時系列データのクラスタリングと並行して，Web ページのクラスタリングについても考察をすすめている．これらの結果については，稿を改めて報告したい．

謝辞

本研究に関して，データの提供ならびに大変有益な議論をして頂いた，北海道大学創成科学研究機構・安住薫助手の研究グループに感謝の意を表します．

参考文献

[1] E. Tomita and T. Seki, “An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique”, Proceedings of the 4th International Conference on Discrete Mathematics and Theoretical Computer Science - DMTCS’03, Springer-LNCS 2731, pp. 278 - 289, 2003.

[2] 宇野 毅明, “大規模グラフに対する高速クリーク列挙アルゴリズム”, 電子情報通信学会技術研究報告, Vol. 103, No.31 (COMP2003 1-8), pp. 55 - 62, 2003.

[3] Y. Okubo and M. Haraguchi, “Creating Abstract Concepts for Classification by Finding Top-N Maximal Weighted Cliques”, Proceedings of the 6th International Conference on Discovery Science - DS’03, Springer-LNAI 2843, pp. 418 - 425, 2003.