4B-04

グラフマイニングでの多重検定補正

杉山 麿人 †‡　Felipe Llinares López§　Niklas Kasenburg¶　Karsten M. Borgwardt§
† 大阪大学産業科学研究所　‡ 独立行政法人科学技術振興機構, さきがけ　§D-BSSE, ETH Zürich
¶DIKU, University of Copenhagen

1　はじめに

グラフ構造をもつデータから知識発見をおこなう**グラフマイニング**（graph mining）は，データマイニングにおける重要な研究分野であり，創薬やタンパク構造の解析などに応用されている．特に，グラフ集合が幾つかのクラスにあらかじめ分類されているとき，あるクラスにのみ頻出している部分グラフは，そのクラスにとって重要な部分構造と考えられる．そして，様々な応用領域において（例えば創薬 [5]），統計的に有意に頻出している部分グラフの発見が求められている．

仮説空間（全部分グラフからなる集合）全体にわたって偽陽性が生じる確率（FWER）を適切に制御するために，**多重検定補正**（multiple testing correction）が必要となる．しかし，以下の計算量的，及び統計的な問題がある．(1) どのようにして数百〜数千万個にもなる部分グラフを全て検定するのか．(2) どうやって FWER を制御するのか．後者の解決のための一般的な多重検定補正は Bonferroni 補正 [1] であり，FWER の上限 α を総検定数 m で割った値 α/m を各検定で用いる．しかし，部分グラフが大量に存在するため，補正後の有意水準が低くなりすぎてしまい，検出力が下がって有意な部分グラフを見落としてしまう危険性がある．さらに，Bonferroni 補正は総検定数 m を必要とし，これは全部分グラフの総数に対応するため，この計算は計算量的に困難である．

寺田ら [7] は，仮説の**検定可能性**（testability）を用いることで，アイテム集合マイニングにおいてこの多重検定問題を解決した．検定可能性とは，もともと Tarone [6] によって導入された概念で，検定可能でない仮説は取り除いてしまっても FWER が変化しない，という性質である．そこで本論文では，この検定可能性をグラフマイニングに適用する．そして，検定可能性と頻出部分グラフマイニングアルゴリズムを組み合わせることで，検定すべき部分グラフを大量に削減することができ，計算の高速化及び検出力の向上が実現できることを報告する．

Multiple Testing Correction in Graph Mining
Mahito SUGIYAMA†‡
†ISIR, Osaka University
‡JST, PRESTO

2　検定可能性

2 つのグラフ集合 $\mathcal{G}$ と $\mathcal{G}'$ を考える．各集合におけるグラフ数を $|\mathcal{G}| = n$ 及び $|\mathcal{G}'| = n'$ とし，$n \leq n'$ と仮定する．グラフ H に対して，$\mathcal{G}$ 中で H を部分グラフとして持つグラフの個数を x とする．すなわち，$x = |\{ G \in \mathcal{G} \mid H \sqsubseteq G \}|$ である．同様に，$\mathcal{G}'$ における出現数を x' とする．このとき，このグラフ H の統計的有意性，すなわち P 値は，x, n, n' からフィッシャーの正確確率検定によって求まる．

グラフ H に対して，$\mathcal{G}$ と $\mathcal{G}'$ 全体における出現数を $f(H) = |\{G \in \mathcal{G} \cup \mathcal{G}' \mid H \sqsubseteq G\}|$ と定義し，$f(H) \leq n$ と仮定する．すると，$f(H), n, n'$ を固定したとき，このグラフ H の P 値が取りうる最小値を $\psi(f(H)) = \psi \circ f(H)$ と書くと，その値は

$$\psi \circ f(H) = \binom{n}{f(H)} \Big/ \binom{n+n'}{f(H)}$$

で与えられる．これは $x = f(H)$ のときの P 値に対応する．したがって，もしこの最小 P 値が有意水準より大きければ，その部分グラフ H は各集合での出現数 x, x' に関わらず，必ず有意にならないことがわかる．

Tarone [6] は，これらの仮説を取り除いてしまっても FWER が維持できることを示した．まず，仮説空間（すべての部分グラフ）$\mathcal{H} = \{H \sqsubseteq G \mid G \in \mathcal{G} \cup \mathcal{G}'\}$ とし，自然数 k に対して $m(k) = |\{ H \in \mathcal{H} \mid \psi \circ f(H) \leq \alpha/k \}|$ と定義する．すると，

$$\begin{aligned} \mathrm{FWER} &\leq \sum \{ \psi \circ f(H) \mid \psi \circ f(H) \leq \alpha/k,\ H \in \mathcal{H} \} \\ &\leq m(k)\frac{\alpha}{k} \leq \alpha \end{aligned}$$

となる．したがって，FWER を α 以下に維持しつつ検出力を最大化する（検定すべき部分グラフ数を最小化する）には，$m(k) < k$ を満たす最小の k を求めればよい．関数 $m(k)$ は単調減少なので，$m(k-1) > k-1$, $m(k) \leq k$ を満たす k を k_{rt} と書くと，これが解である．このとき

$$\tau(\mathcal{H}) = \{H \in \mathcal{H} \mid \psi \circ f(H) \leq \alpha/k_{\mathrm{rt}}\}$$

の各要素を**検定可能な部分グラフ**（testable subgraph）と呼ぶ．検定可能な部分グラフのみを検定すればよく，各検定で用いる補正後の有意水準は $\alpha/|\tau(\mathcal{H})|$ となり，Bonferroni 補正と比べると，検出力は必ず高くなる．

3　頻出部分グラフマイニングを用いた列挙

頻出部分グラフマイニングの手法を用いることで，効率的に検定可能部分グラフ集合 $\tau(\mathcal{H})$ を列挙する．頻出部分グラフマイニングでは，出現数が σ より多い全ての部分グラフを列挙する．ここで ψ は単調減少関数なので [7, Supporting Text 4]，全ての頻出部分グラフ H に対して $\psi \circ f(H) \leq \psi(\sigma)$ が成り立つ．したがって，条件

$$|\left\{ H \in \mathcal{H} \,\middle|\, f(H) \geq (\sigma_{\mathrm{rt}} - 1) \right\}| > \alpha/\psi(\sigma_{\mathrm{rt}} - 1),$$
$$|\left\{ H \in \mathcal{H} \,\middle|\, f(H) \geq \sigma_{\mathrm{rt}} \right\}| \leq \alpha/\psi(\sigma_{\mathrm{rt}})$$

を満たす出現数 σ_{rt} に対して，$k_{\mathrm{rt}} = \alpha/\sigma_{\mathrm{rt}}$ から，$\tau(\mathcal{H})$ は閾値 σ_{rt} 以上の頻出部分グラフと一致する．すなわち，この閾値 σ_{rt} を求めて，頻出部分グラフを列挙すればよい．以下では，2 つの戦略を述べる．

減少探索法． まず，寺田ら [7] によって提案された手法 LAMP で用いられている，減少探索法について述べる．最初に，出現数を最大値 $\sigma = n$ に設定し，頻出部分グラフマイニング手法を適用し，そのときの部分グラフ総数 m を計算する．そして，条件 $m > \alpha/\psi(\sigma)$ が満たされるまで，σ を一つづつ減少させ，頻出部分グラフマイニング手法の適用による総数 m の計算を繰り返す．もし満たされたら，$\sigma_{\mathrm{rt}} = \sigma + 1$ が求める閾値である．

増加探索法． 上の戦略とは逆に，σ を増加させる方法である．最初に，$\sigma = 1$ に設定し，頻出部分グラフマイニング手法を実行する．その実行中，頻出部分グラフを見つける度に，そこまでの頻出部分グラフの総数 m に対して条件 $m > \alpha/\psi(\sigma)$ を確認し，もし満たされたら打ち切る．そして，σ を 1 増加し，また頻出部分グラフマイニング手法を実行する．これを繰り返し，もし打ち切られず頻出部分グラフマイニングが最後まで終了したら，その時の σ が求める閾値である．これは，著者らによって提案され [4]，また同時に，湊らによってアイテム集合発見で同様の手法が提案されている [2]．

4　実験

実データ実験によって，Bonferroni 補正を用いた総当り法，すなわち，全部分グラフを列挙する手法と，検定可能な部分グラフのみを列挙する手法（減少探索法・増加探索法）を比較する．頻出グラフマイニング手法として，最も高速な手法として知られている Gaston [3] を用いる．実験を通して有意水準 $\alpha = 0.05$ とする．ベンチマークとしてよく用いられている 3 つのグラフデータ PTC (MR) , D&D, NCI41 を用いる（表 1参照）．

結果を図 1に示す．検定可能性を用いることで，大幅に検定数を減らすことに成功している．これは，FWER を $\alpha = 0.05$ 以下に保ちつつ，検出力が増加していることを意味する．また，実行時間については，増加探索法が最も速く，総当り法と比べると多くの場合で 100 倍以上高速である．また，増加探索法よりも減少探索法が遅いが，これは，最終的な閾値 σ_{rt} が通常小さく（20 程度），減少探索法では何度も頻出部分グラフマイニング手法を繰り返さなければならないためと考えられる．

表1　グラフデータ．

データ	$\lvert\mathcal{G} \cup \mathcal{G}'\rvert$	$\lvert\mathcal{G}\rvert$	avg.$\lvert V\rvert$	avg.$\lvert E\rvert$	平均次数
PTC (MR)	584	181	31.96	32.71	2.01
D&D	1178	691	284.32	715.66	4.98
NCI41	27965	1623	47.97	50.15	2.09

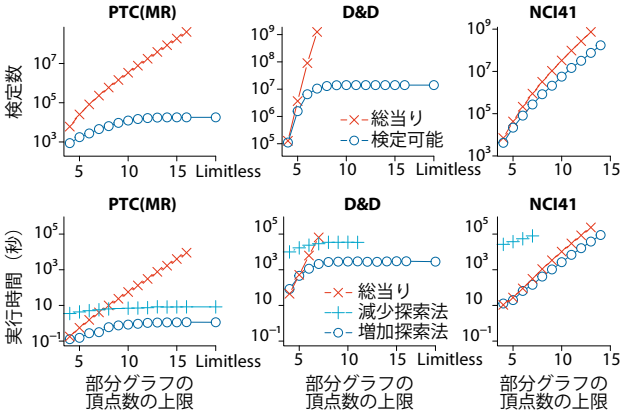

図1　検定数（上段）と実行時間（下段）．

5　まとめ

本論文では，2 つのグラフ集合において，片方のクラスのみに統計的に有意に出現している部分グラフをすべて列挙する手法を提案した．検定可能性を用いることで，指数関数的に膨らむ探索空間の削減，および，多重検定補正による偽陽性が生じる確率の制御をともに実現した．

参考文献

[1] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[2] S. Minato, T. Uno, K. Tsuda, A. Terada, and J. Sese. A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration. In *ECMLPKDD*, pages 422–436, 2014.

[3] S. Nijssen and J. N. Kok. A quickstart in frequent structure mining can make a difference. In *KDD*, pages 647–652, 2004.

[4] M. Sugiyama, F. L. López, N. Kasenburg, and K. M. Borgwardt. Multiple testing correction in graph mining. *arXiv:1407.0316*, 2014. (accepted to SDM15).

[5] I. Takigawa and H. Mamitsuka. Graph mining: Procedure, application to drug discovery and recent advances. *Drug Discovery Today*, 18(1–2):50–57, 2013.

[6] R. E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, 46(2):515–522, 1990.

[7] A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese. Statistical significance of combinatorial regulations. *Proc. Natl. Acad. Sci. USA*, 110(32):12996–13001, 2013.