様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ２６ 日現在

機関番号：１４４０１
研究種目：研究活動スタート支援
研究期間：2014～2015
課題番号：２６８８００１３

研究課題名（和文）グラフ構造データから統計的に有意に頻出する部分構造を発見する手法の研究

研究課題名（英文）Data Mining Methods for Discovering Statistically Significant Substructures of Graphs

研究代表者
杉山　麿人（SUGIYAMA, Mahito）

大阪大学・産業科学研究所・助教

研究者番号：１０７３３８７６

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：本研究では，グラフ構造を持つデータの集合から，統計的に有意に頻出するグラフの部分構造を効率的に発見する手法を構築した．この手法を用いることで，例えば創薬において，目的の効果を持つ化合物群が共通して持っている部分構造を検出することができる．構築した手法は，偽陽性と呼ばれる，本来は効果を持たないのに効果を持っていると誤ってみなされてしまう部分構造の割合を制御する．これによって，創薬や生体機構の解析，マーケティングなどの応用先の領域に対して，より信頼性の高いマイニング結果を提供することができる．

研究成果の概要（英文）：We have developed significant subgraph mining methods, which find substructures of graphs that are statistically significantly enriched in a class of transactions. In drug discovery, for example, the methods enable us to find substructures of chemical compounds which are significantly associated with a particular activity such as an anticancer activity. Our methods control the false positive rate, the probability of erroneous discoveries that actually do not have a particular effect, which is essential to provide reliable results in a number of application domains from drug discovery to social network analysis.

研究分野：機械学習・データマイニング

キーワード：グラフマイニング　パターンマイニング　仮説検定　多重検定　統計的有意性

１．研究開始当初の背景

グラフ構造を持つデータが実世界の様々な分野で獲得され，大量に蓄積されてきている. 例えば, 化合物のデータが PubChem に，タンパク質のデータが Protein Data Bank (PDB) に，遺伝子ネットワークのデータが KEGG に，そして，ソーシャルネットワークのデータが Web 上に広く蓄積されている．これら大量のグラフ構造データからの知識発見を目的とするグラフマイニングは，データマイニングにおける主要なトピックの 1 つとして盛んに研究され，創薬，生体機構の解析，マーケティングなどの幅広い分野へ応用されている．

グラフマイニングにおける最も基本的な問題が，与えられた多数のグラフの中で頻出している部分グラフを全て列挙する問題である．これら頻出部分グラフは，多くのグラフに共通して現れている重要な性質であるとみなされ，応用先の領域で更なる解析の対象となる．ところが実際には，多くの場合，単なる頻度ではなくそれら部分グラフが統計的に有意に頻出しているかどうかが問題となる [1]. 例えば，創薬においては，目的の治癒効果をもつ化合物（グラフ）の集合のなかで統計的に有意に頻出している部分構造（部分グラフ）に興味があり，生体機構の解析においては，目的の結合性質をもつタンパク質集合のなかで統計的に有意に頻出している分子構造に興味がある．

このように，統計的に有意に頻出する部分グラフの発見は，数理的にも基本的な問題であり，かつ様々な応用領域で重要であるにも関わらず，国内・国外問わずこれまで研究がほとんど進んでいなかった．特に，この問題を解決するためには，以下の 2 つの本質的な課題を同時に解決する必要がある．

課題 1. 計算量の爆発：
部分グラフの個数が，グラフの持つ組合せ的な性質によって指数関数的に増大する．さらに，各部分グラフについて，それが他のグラフに含まれているかどうかの判定を繰り返しおこなう必要がある．この判定は NP 完全問題として知られており，その結果，グラフのサイズが大きくなると計算量が爆発して部分グラフの列挙が不可能になる．

課題 2. 多重検定に起因する偽陽性の増加:
統計的な有意性を調べるためには，各部分グラフを検定する必要がある．しかし，その総数が指数関数的に増大する（課題 1）ため，多重検定補正を行わないと，どこかの部分グラフで偶然有意と判定される（偽陽性となる）確率（family-wise error rate; FWER）が増大し，その結果，多くの偽陽性部分グラフが発生してしまう．

グラフマイニングにおいては，これまで課題 1 が主な研究対象であり, gSpan [4] を始めとして数多くの効率的なアルゴリズムが提案されてきた．ところが，課題 2 を対象とする研究はこれまで存在しなかった．例えば，Yan ら [3] は統計的に有意な部分グラフを検出する手法を提案しているが，多重検定補正は考慮されておらず，偽陽性部分グラフが多く含まれてしまう危険性がある．

一方，統計学では，現在でもよく用いられる Bonferroni 法をはじめとして多くの多重検定補正の手法が提案されてきた．ところが，ほとんどの既存の補正法は，検定の総数，すなわち部分グラフの総数を要求する．これは，出現数が 1 以上の全ての頻出部分グラフを求めることに対応し，最先端の効率的なアルゴリズムを用いても困難である．また，その莫大な検定数によって，補正をおこなうと FWER の維持と引き換えに検出力が低下し過ぎてしまい，有意な部分グラフを発見できなくなってしまう（偽陰性が増える）危険性も存在する．

近年，寺田ら [2] が頻出アイテム集合発見と多重検定補正を組み合わせることで遺伝子発現に関わる転写因子の同定に成功しており，本研究では，そこでのアプローチをグラフマイニングへ応用することで上記 2 つの課題を解決することを目指した．

[1] Takigawa, I. and Mamitsuka, H.: Graph mining: procedure, application to drug discovery and recent advances, *Drug Discovery Today*, **18**(1–2), 50–57 (2013).
[2] Terada, A., Okada-Hatakeyama, M., Tsuda, K., Sese, J.: Statistical significance of combinatorial regulations, *Proc. Natl. Acad. Sci. USA* (2013).
[3] Yan, X., Cheng, H., Han, J., and Yu, P.~S.: Mining Significant Graph Patterns by Leap Search, In *Proc. 2008 ACM SIGMOD International Conference on Management of Data*, 433–444 (2008).
[4] Yan, X. and Han, J.: gSpan: Graph-Based Substructure Pattern Mining, In *Proc. 2002 IEEE International Conference on Data Mining*, 721–724 (2002).

２．研究の目的

本研究の目的は，与えられたグラフ集合から，統計的に有意に頻出する部分グラフをすべて発見する手法を構築することである．

z この目的を達成するため，部分グラフの全列挙を避けて計算量と検定数をともに抑えつつ，適切に多重検定補正をおこなうことで，先に述べた 2 つの課題（計算量爆発・多重検定）を同時に解決する．より具体的には，以下の 3 項目に取り組んだ．

- グラフマイニングでの計算量爆発を避けることのできる，検定（部分グラフ）総数を必要としない多重検定補正法の設計.
- その補正法を組み込んだ, 有意な部分グラフを効率的に列挙するアルゴリズムの構築と実装.
- 実世界のグラフ構造データへの適用による，実装したアルゴリズムの実効性検証.

３．研究の方法

部分グラフの総数を数えることなく多重検補正をおこなうために，検定可能性，実効的検定数，そして，Westfall と Young によるランダム置換多重検定法という 3 つの多重検定法を導入する．さらに，それらと既存の頻出部分グラフ発見アルゴリズムを融合することで，多重検定補正を達成しつつ計算量爆発を避けて効率的に有意な部分グラフを列挙するアルゴリズムを構築する．

これらの多重検定法は独立に用いることができるため，本研究においては，まず 1 年目で検定可能性及び実効的検定数を用いた手法を構築し，次に 2 年目で Westfall と Young のランダム置換法を用いる．これらの手法を用いることで，本研究を通して，以下の問題設定に取り組み，その解決を目指す．

問題：多数のグラフが与えられ，各グラフは 2 つのクラスのどちらかに分類されている．このとき，一方のクラスにおいて他方と比べて有意に頻出している部分グラフを全て発見する．

以下では，各年度における研究方法を詳細に述べる．

(1) 平成 26 年度

上記の問題設定において，各部分グラフの出現状況は 2x2 分割表として記述されるため，各部分グラフの有意性は，標準的な検定法であるフィッシャーの正確確率検定によって検定できる．最もよく使われている多重検定補正法は，Bonferroni 法である．各検定における有意水準を α とすると，まず部分グラフの総数 m を求めて，各部分グラフに対して有意水準 α / m で検定をおこなう．しかし，m を求めるために全ての部分グラフを列挙する必要があるため，その計算が困難である．さらに，その莫大な m によって有意性の判定が保守的になりすぎ，検出力が過度に下がってしまう．

本研究では，Tarone [2] が提案した仮説の検定可能性を用いてこれらの問題を解決する．各仮説に対して，P 値の取りうる下限値が有意水準よりも小さければ，この仮説は検定可能という．Tarone は，Bonferroni 法において検定可能でない（P 値の下限値が有意水準より大きい）仮説をあらかじめ取り除いても family-wise error rate (FWER) が厳格に維持できることを示した．すなわち，検定可能な仮説の数を m' とすると，それらを有意水準 α / m' で検定することができ，m' が m より小さくなればなるほど検出力が上がる．この Tarone の結果は，本研究での部分グラフの検定に直接適用することができ，部分グラフの総数を計算することなく検定可能な部分グラフのみを列挙すればよい．

さらに，頻出部分グラフと検定可能な部分グラフには密接な関係がある．すなわち，部分グラフの頻度とその検定可能性に相関関係があり，検定可能性（P 値の下限）はその部分グラフのグラフ集合全体における頻度から決定される．この性質を利用することで，既存の頻出部分グラフ発見アルゴリズムに基づいた検定可能部分グラフの列挙アルゴリズムを構築する．

さらに，ここまでで構築した手法に実効的検定数 [1] を導入することで，更に検定すべき部分グラフ数を減らして検出力を上げる．実効的検定数とは統計遺伝学において提案された手法であり，仮説間に相関関係があるとき，Bonferroni 法での多重検定補正 α / m で用いる総検定数 m をこの実効的検定数に減らしても FWER が維持できる（各検定間が独立ならば実効的検定数は m と等しくなる）．本研究では，検定可能な部分グラフのみから実効的検定数を計算することによって，計算を効率化する．

アルゴリズムを実装した後，グラフ分類の研究などにおいて標準的に用いられている化合物などの実世界のベンチマークデータ（NCI データセットなど）を用いて，その有効性を検証する．

(2) 平成 27 年度

ここまでで構築した手法における FWER の制御は，常に保守的となり，実際の値は設定した閾値 α よりも大きくなる．FWER= α を達成するために，Westfall と Young が提案したランダム置換多重検定法を部分グラフマイニングのアルゴリズムに統合する．あらかじめランダム置換したクラスラベルを用意しておくことで，部分グラフマイニングの各時点で最終的な FWER を推定し，計算を効率化する．

前年度と同様に，アルゴリズムを構築・実装し，実データを用いて有効性を検証する．

[1] Nyholt, D. R.: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other, *The American Journal of Human Genetics*, **74**(4), 765–769 (2004).

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

４．研究成果

検定可能性，実効的検定数，そして，Westfall と Young のランダム置換多重検定法を部分グラフマイニングのアルゴリズムに導入することに成功した．これによって，FWER（偽陽性の割合）を制御しつつ，統計的に有意に出現する部分グラフをすべて発見することが初めて可能となった．

鍵となるのは，どの多重検定法においても，既存の頻出部分グラフ発見アルゴリズムによって部分グラフを列挙していく各時点で，最終的な偽陽性割合が推定できる，という理論的成果である．これによって，既存の頻出部分グラフ発見アルゴリズムの効率性を損なうことなく，高速かつ省メモリに統計的に有意な部分グラフが発見できるようになった．これは，グラフマイニングだけでなく，

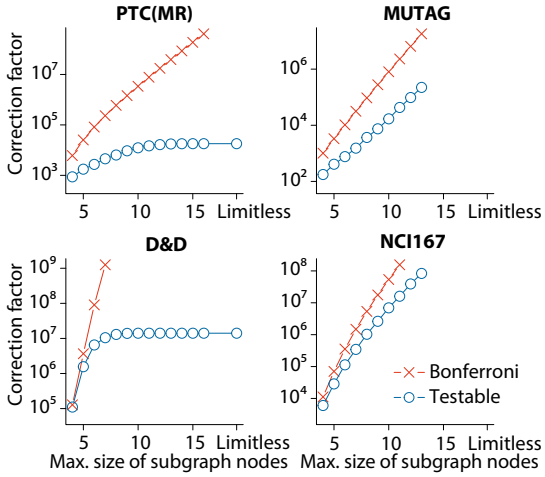

図 1: 検定すべき部分グラフの数．部分グラフの総数（赤）と検定可能部分グラフ数（青）．

アイテム集合や系列，木パターンなど，パターンマイニング一般において用いることのできる手法であるため，その汎用性は高い．

化合物やタンパク質などの，部分グラフマイニングでよく用いられる実世界のベンチマークデータを用いて構築した手法の効率性を検証した．その結果，検定すべき部分グラフの総数を大幅に抑えることができることを確認した（図 1）．これによって，既存手法より 100 倍から 1000 倍程度の高速化を達成し，かつ偽陽性の割合が適切に制御されていることを確認した．

研究成果は，データマイニングのトップ会議 KDD や主要会議 SDM に採録され，さらに，遺伝子データへ応用した研究はバイオインフォマティクスのトップ会議である ISMB に採録されるなど，世界的にも評価を得ている．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件，全て査読あり）

① Sugiyama, M., Nakahara, H., Tsuda, K.: Information Decomposition on Structured Space, *IEEE ISIT* 2016 (to appear).
② Sugiyama, M., Borgwardt, K. M.: Halting in Random Walk Kernels, *Advances in Neural Information Processing Systems* 28, 1630–1638, 2015.
③ Llinares-López, F., Sugiyama, M., Papaxanthos, L., Borgwardt, K. M.: Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing, *Proc. of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 725–734, 2015, DOI: 10.1145/2783258.2783363
④ Llinares-López, F., Grimm, D. G., Bodenham, D. A., Gieraths, U., Sugiyama, M., Rowan, B., Borgwardt, K. M.: Genome-Wide Detection of Intervals of Genetic Heterogeneity Associated with Complex Traits, *Bioinformatics*, **31**(12), i240–i249, 2015, DOI: 10.1093/bioinformatics/btv263
⑤ Sugiyama, M., Llinares-López, F., Kasenburg, N., Borgwardt, K. M.: Significant Subgraph Mining with Multiple Testing Correction, *Proc. of the 2015 SIAM International Conference on Data Mining*, 37–45, 2015, DOI: 10.1137/1.9781611974010.5
⑥ Sugiyama, M., Otaki, K.: Detecting Anomalous Subgraphs on Attributed Graphs via Parametric Flow, *New Frontiers in Artificial Intelligence*, LNCS **9067**, 340–355, 2015, DOI: 10.1007/978-3-662-48119-6_26

〔学会発表〕（計 5 件）

① Sugiyama, M.: Statistical Analysis on Order Structures, 3rd mini-symposium on Computations, Brains and Machines, 2016 年 3 月 17 日，理研 BSI（埼玉県・和光市）
② 杉山 麿人, Borgwardt, K.M.: ランダムウォークグラフカーネルの停止に関する解析，第 18 回情報論的学習理論ワークショップ（IBIS2015），2015 年 11 月 25–28 日，つくば国際会議場（茨城県・つくば市）
③ 杉山 麿人: 統計的パターンマイニング，第 14 回情報科学技術フォーラム，2015 年 9 月 15–17 日，愛媛大学（愛媛県・松山市）
④ 馬場 祥人，杉山 麿人，鷲尾 隆: サンプリングを用いた高速頻出パターンマイニング，第 29 回人工知能学会全国大会，2015 年 5 月 30 日–6 月 2 日，公立はこだて未来大学（北海道・函館市）
⑤ Sugiyama, M.: Multiple Testing Correction in Graph Mining, Tokyo Workshop on Statistically Sound Data Mining, 2015 年 2 月 16 日，産総研 臨海副都心センター（東京都・江東区）

〔ホームページ〕
http://mahito.info

６．研究組織
(1)研究代表者
杉山 麿人（SUGIYAMA, Mahito）
大阪大学・産業科学研究所・助教
研究者番号：10733876

(2)研究分担者
該当なし

(3)連携研究者
該当なし