研　究　報　告　書

「統計的有意性を担保する超高速パターン発見技術の創出」

研究タイプ：通常型
研究期間：　平成 26 年 10 月～平成 30 年 3 月
研　究　者：　杉山　麿人

1.　研究のねらい

ビッグデータからの知識発見を目的とするデータマイニングは，化学や生物学などの基礎科学から経営やマーケティングへの応用に至るまで，幅広い分野で活用されている．特に，データベースからの頻出パターン発見は，データマイニングにおける最も基本的な問題の一つであり，データに隠れた有益な規則を見つけるための手法として盛んに研究されてきた．パターンとは，同時に出現する対象の組合せであり，例えば，一緒に売れた商品や，文書中で共起して現れる単語（アイテム集合），特定の機能をもつ化合物が共有している構造（部分グラフ）などが，解析の対象となる．

しかし，多くの応用分野において求められているのは，単に頻出しているパターンではなく，統計的に有意に頻出しているパターン，すなわち，偶然に発生すると期待されるよりも高い頻度で出現するパターンである．このようなパターンを発見することで，例えば創薬において，化合物データベースから目的の活性に統計的に有意に関連している化合物の部分構造を見つけることができ，遺伝学においては，ゲノムデータベースから対象の疾患に統計的に有意に関連している遺伝子変異の組み合わせを発見することができる．さらに，統計的有意性を用いることで，偽陽性パターンの割合，すなわち，関連があるとして発見されてしまったが実際には関連がないパターンの割合を，任意の値に制御することができる．

統計的有意性を担保したパターンの発見は，応用分野において必須の要請があり，かつ数理的にも基礎的な問題であるにも関わらず，研究が進んでいなかった．そこで本研究では，この問題を解決し，大規模データベースから統計的に有意に頻出しているパターンを直接，高速に列挙する統計的パターンマイニング（significant pattern mining）のアルゴリズム群を構築し，その妥当性を担保するための統計理論を整備し体系化することをねらいとした．

2.　研究成果

（1）概要

本研究では，一貫してデータからの統計的データマイニング・機械学習技術の開発に関して研究を進め，主に 3 つの研究テーマに関して，それぞれ成果を挙げた．

研究テーマA「統計的パターンマイニング技術の確立」では，統計的に有意に出現するパターンを発見する統計的パターンマイニングの技術を確立した．特に，グラフ構造をもつデータから，統計的に有意に現れる部分グラフを発見しつつ，全候補グラフにわたって偽陽性の割合を制御することができる手法を初めて構築することに成功した．さらに，この手法を含むより一般的な統計的パターンマイニング技術の高速化・省メモリ化を実現し，state-of-the-art を確立した．構築した手法を遺伝子データへ適用することによって，これまで知られていなかった新たなパターン（塩基対の組合せ）が発見できることを報告した．

研究テーマB「情報幾何を用いた階層的な確率モデルの解析」では，統計的パターンマイニングで扱う階層的な空間を対数線形モデルによって確率モデル化し，統計的及び情報理論的構造を詳細に解析することで，情報幾何とパターンマイニングの密接な関係を理論的に明らかにした．その得られた結果を応用することで，行列バランシングという行列の一種の正規化が高速に解けることを示し，その一般化であるテンソルバランシングの実現に成功した．既存の行列バランシング手法と比較して，10,000 倍以上の高速化を実現した．
研究テーマC「グラフ構造データに対する機械学習技術の開発」では，グラフ構造を持つデータに対する機械学習手法の解析及び構築をおこなった．特に，グラフ間の類似度を測るグラフカーネル手法の解析をおこない，既存のベースライン手法として知られている幾何ランダムウォークが適切な手法ではないことを明らかにした．さらに，グラフカーネル手法を網羅した R 及び Python で利用可能なパッケージ graphkernels を公開した．

（2）詳細

研究テーマA「統計的パターンマイニング技術の確立」

統計的パターンマイニング（significant pattern mining）と呼ばれる，データベースから統計的に有意に出現するパターンを発見する技術の確立を目的とし，研究を進めた．この技術を達成するためには，以下の 2 つの課題を解決する必要がある．

- 計算論的な課題（計算量の爆発）：データベースのサイズが大きくなると，パターンの探索空間が組合せ爆発を起こし，パターンの探索・列挙が困難になる．
- 統計的な課題（多重検定に起因する偽陽性の増加）：パターン総数が指数関数的に増大するため，各パターンの検定において多重検定補正を行わないと，偶然有意と判定される（偽陽性となる）確率が増大し，大量の偽陽性パターンが発生してしまう．

これら 2 つの課題を，Tarone の検定可能性と Apriori 法を組み合わせることで解決した．Tarone の検定可能性によって確実に有意ならない不必要なパターンを同定し，Apriori 法によってそれら不必要なパターンを効率的に枝刈りすることで，2 つの課題を同時に解決した．

本研究では，まず統計的部分グラフマイニング（図 1）を達成した．ラベル付けされたグラフのデータベースから，特定のクラスにおいて他のクラスと比べて統計的に有意に頻出している部分グラフを全て発見し，かつ全体における偽陽性の割合を適切に制御することに成功した．Tarone の検定可能性を導入して検定可能でない部分グラフを同定し，Apriori 法と組み合わせることでそれらを効率的に除去する手法を構築し，計算機上で実装した．実データを用いた検証の結果，既存手法とくらべて 1000 倍程度の高速化を達成しつつ，偽陽性の割合を適切に制御できることを示した[論文 1]．

さらに，ランダム置換を用いた統計的多重検定法を組み込むことで，より正確に偽陽性の割合を制御できる手法 Westfall-Young light の構築に成功した[論文 2]．部分グラフや組合せ集合など様々な対象に適用可能なパターン列挙アルゴリズムの構築・実装をおこない，実世界でのベンチマークデータによる性能の検証によって，既存手法よりも高速かつ省メモリで統計的パターンマイニングが達成できることを示した．この手法は，現在でも統計的パターンマイニングの state-of-the-art である．

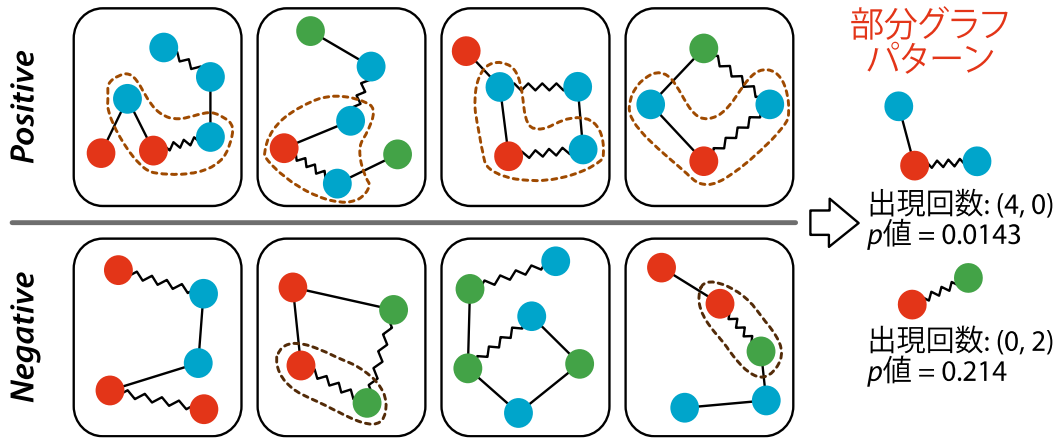

図1：統計的部分グラフマイニング．P 値が多重検定補正されたしきい値より小さい部分グラフパターンをすべて見つける．

研究テーマB「情報幾何を用いた階層的な確率モデルの解析」

パターンマイニングで現れるパターンの集合がなす空間は，半順序構造と呼ばれる階層的な構造を必ず持つ．半順序構造とは，有向非巡回グラフ（DAG）と等価な構造であり，集合の包含関係をはじめとして幅広い対象が半順序構造を保有しており，計算機科学における本質的な離散構造・階層構造である．そこで，半順序構造に対する確率モデル（対数線形モデル）を導入することで，パターン空間が持つ統計的性質や情報理論的性質をより詳細に解析することを目的として，研究を進めた．

結果として，この半順序構造上の対数線形モデルによって生成される確率分布族が，情報幾何で知られている双対平坦多様体となることを発見した．これは，パターン空間の確率分布が指数型分布族に含まれることを意味し，パターンの頻度は十分統計量に対応する．さらに，確率モデルの学習が最尤法に代表される部分多様体への射影として実現できる（図 2）．これらの強力な性質は，本研究によって初めて明らかになった．

さらに，この多様体の構造を利用した数値計算アルゴリズムを設計し，行列やテンソルのバランス化に適用することで，従来法よりも高速にバランス化を達成するアルゴリズムを構築した．行列のバランス化とは，各列，各行に対する定数倍のみを用いて各列，各行の和がどれも 1 となるようにする操作であり，経済学における産業連関表の解析や，生物学での Hi-C データ解析において標準的処理として用いられている．提案手法は，初めて行列バランス化の一般化であるテンソルバランス化を達成し，また行列バランシングにおいては既存手法より 10,000 倍程度の高速化を達成した（論文 3）．

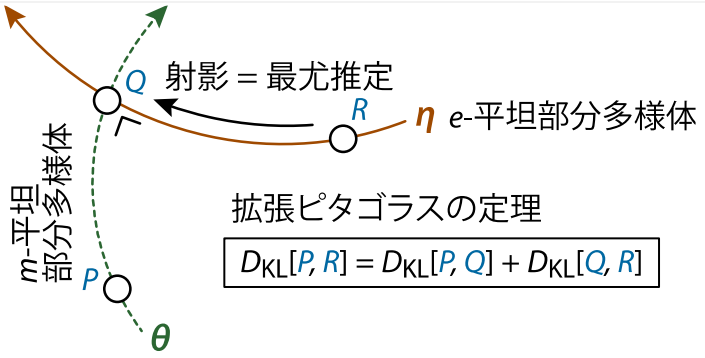

図 2: パターン空間の対数線形モデルによって定まる双対平坦多様体．θ は対数線形モデルの係数，η はパターンの頻度に対応し，これらは必ず直交する．

研究テーマC「グラフ構造データに対する機械学習技術の開発」

グラフ構造を持つデータを解析するための機械学習技術の研究をおこなった．特に，グラフカーネルと呼ばれる，グラフ間の類似度を測る手法を解析し，標準的なベースライン手法として知られている幾何ランダムウォークカーネルがベースラインとして適切ではなく，k ステップランダムウォークカーネルが適していることを，理論と実験両面から明らかにした[論文 4]．

さらに，グラフカーネルの主要な手法を網羅した R 及び Python で利用可能なパッケージ graphkernels を公開した（図 3）[論文 5]．本報告書執筆時点で，既に 10,000 回以上のダウンロードがあり，グラフカーネルを利用したデータ解析をおこなう際の標準的パッケージとして定着しつつある．

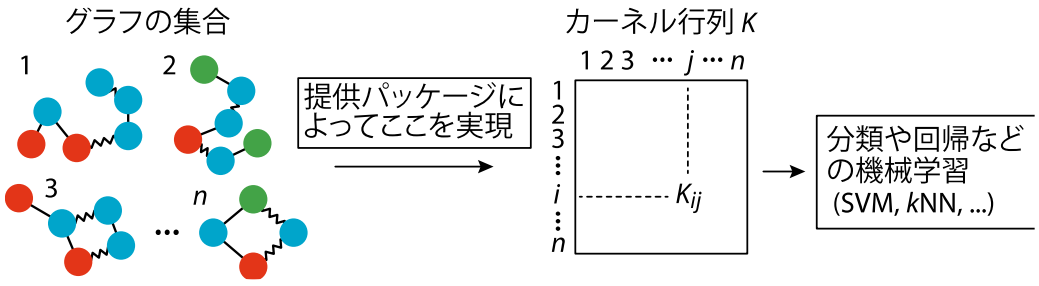

図 3: パッケージ graphkernels の概要．グラフの集合を入力とし，カーネル行列を出力する．

3. 今後の展開

主に以下の 2 つの研究トピックに関して，研究を展開する．

研究トピックA「階層的構造に対する対数線形モデルを用いた機械学習の理論的解析」

深層学習で用いられている階層的なモデルを，本研究で提案した階層的構造に対する対数線形モデルと，その情報幾何的性質を用いて解析する．特に，ボルツマンマシンなどの生成モデルを対象とする．

研究トピックB「統計的パターンマイニングなどを用いた応用研究」

統計的パターンマイニングには幅広い応用があるが，それらは未だ発展途上である．そこで，脳活動データや医療データなどに適用することで，新規の科学的発見を目指す．さらに，本研究で

提案した対数線形モデルを用いることで，変数間の高次相関が持つ情報量を分解し，取り出すことができる．そこで，上記データへの適用可能性を探る．

4．　評価

（1）自己評価

研究目的の達成状況

申請時に掲げた統計的パターンマイニング技術の確立，という研究目的を，計画通り達成することができた．さらに，この研究を進めることで，申請時にはまったく無かった，情報幾何を用いた階層的な確率モデルの解析及び，グラフ構造データに対する機械学習技術の開発という新たな 2 つの研究テーマについて研究を進めて，それぞれ成果を挙げることができた．したがって，本研究の研究目的は，十分に達成できた．

研究の進め方（研究実施体制及び研究費執行状況）

予定通り研究を進め，各年度で研究費を計画通り執行することができた．

研究成果の科学技術及び社会・経済への波及効果（今後の見込みを含む）

現時点では学術的な貢献に留まっているが，本研究で達成した研究成果は，独創的かつインパクトのある内容であると自負している．したがって，今後応用研究に取り組むことで，よりインパクトの大きい社会実装へと繋げることができると考えている．

申請時の目的をすべて達成した．さらに，申請時には明らかでなった課題を解決した．

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータ解析技術を用いて、科学の世界で認められる発見を行うためには、その解析結果が統計的に有意でなければならない。本研究の第一の成果は、偽陽性の割合を制御して統計的に有意な頻出パターンを見出す方式の確立と高速で省メモリなアルゴリズムの提案である。第二の成果は、パターンの空間に自然な半順序構造を導入し、その統計的構造を理論的に深く分析することで、情報幾何の概念との関連性を見出したことである。さらに、この関連性を活用して、行列バランス化に関しては既存のものをはるかに凌ぐ高速化を達成し、その一般化であるテンソルバランス化に関しては世界で初めて解法を提案した。そして、第三の成果としては、標準的なグラフカーネルの分析を行うとともに、R と Python から利用可能なグラフカーネルのパッケージを公開した。

それぞれの成果がトップカンファレンスなどで発表されており、優れた学術的成果をあげている。当初の構想にほぼ対応するのは第一の成果であり、第二、第三の成果を得たことで、新たな視界が開けている。今後、情報幾何を用いた理論的な分析をさらに深めるとともに、インパクトのある応用につなげることを期待したい。

5．　主な研究成果リスト

（1）論文（原著論文）発表

1. Sugiyama, M., Llinares-López, F., Kasenburg, N., Borgwardt, K.M.: Significant Subgraph

Mining with Multiple Testing Correction, Proceedings of the 2015 SIAM International Conference on Data Mining (SDM2015), 37–45, 2015

2. Llinares–López, F., Sugiyama, M., Papaxanthos, L., Borgwardt, K.M. Fast and Memory–Efficient Significant Pattern Mining via Permutation Testing, Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2015), 725–734, 2015

3. Sugiyama, M., Nakahara, H., Tsuda, K.: Tensor Balancing on Statistical Manifold, Proceedings of the 34th International Conference on Machine Learning (ICML2017), 70, 3270–3279, 2017

4. Sugiyama, M., Borgwardt, K.M.: Halting in Random Walk Kernels, Advances in Neural Information Processing Systems (NIPS2015), 28, 1630—1638, 2015

5. Sugiyama, M., Ghisu, E., Llinares–López, F., Borgwardt, K.M.: graphkernels: R and Python Packages for Graph Comparison, Bioinformatics, btx602, 2017

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

論文・学会発表

1. 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 (Proceedings of ISMB/ECCB 2015)

2. Sugiyama, M., Nakahara, H., Tsuda, K.: Information Decomposition on Structured Space, Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT), 575–579, 2016

3. Sugiyama, M.: Significant Pattern Mining on Graphs, 10th International Conference on Multiple Comparison Procedures, 2017

4. 杉山 麿人: 統計的有意性を担保するパターンマイニング技術, オペレーションズ・リサーチ誌, 62(4), 2017

受賞

IBISML 研究会賞ファイナリスト, ランダムウォークグラフカーネルの停止に関する解析, 杉山 麿人, Karsten Borgwardt, 2015