様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　９　日現在

機関番号：　１０１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２４２

研究課題名（和文）不均質なグラフ集合に対する教師付き学習系の設計

研究課題名（英文）Supervised learning for inhomogeneous set of graphs

研究代表者

瀧川　一学（Takigawa, Ichigaku）

北海道大学・情報科学研究科・准教授

研究者番号：１０３７４５９７

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：創薬における定量的構造活性相関など、対象データがグラフ表現で与えられる場合の教師付き学習において、データの出自、計測対象の多因子性などの影響により、実データは「不均質さ」を伴う。この問題に対処するため主に４点の研究を行った。１：可能な部分グラフ特徴の有無をすべて考え、その中から予測モデルの学習と必要な特徴の学習を同時に行う手法の開発と分析を行った。２：部分グラフ特徴の空間を0/1値、相関構造、冗長性等の観点から分析した。３：全ての部分グラフ特徴を対象に決定木・回帰木を学習する手法の構築とアンサンブル学習拡張を行った。４：部分グラフ特徴にワイルドカードを許容する場合の手法構築と分析を行った。

研究成果の概要（英文）：When supervised learning over graphs is applied to, for example, real molecular graphs in QSAR, it suffers from the 'inhomogeneity' originated from mixing different data sources and different underlying mechanisms. To address this problem, we conducted research on the following four topics: 1) develop and analyze computational methods for simultaneous learning of predictive model and relevant subgraph features among all possible ones; 2) analyze the properties of feature space of subgraph indicators with real datasets, in particular, boolean structures, correlation structures, and redundancy; 3) develop computational methods for learning decision and regression trees over all possible subgraph features, and its ensemble learning by boosting; 4) develop a relaxed feature representation by introducing wildcard labels to node and edge labels of graphs.

研究分野：　機械学習

キーワード：　機械学習　グラフ　潜在構造

１．研究開始当初の背景

(1) 化学構造式とそれが表現する化合物の生物活性など、多数の構造のデータに基づく統計的モデリングにおいて、対象構造が何らかのグラフ表現で抽象化できる場合の教師付き学習は創薬の定量的構造活性相関用途で研究されてきた。

(2) 各種のアッセイとその化学構造情報の公的リポジトリ PubChem や ChEMBL など実際の大規模な構造データにおいては、データ出自や計測環境の異なるデータが混合された多混合母集団となってしまう。こうした構造的「不均質さ」を伴うデータの扱いへの対処が必要である。

２．研究の目的

(1) 分子構造を表現するグラフ構造データの実際の不均質さを伴うデータを分析し、こうした構造的「不均質さ」を伴うグラフデータに対する教師付き学習の精度向上および特徴空間の性質の理解を目指す。

(2) グラフ構造データの教師付き学習のいくつかの既存アプローチを整理し、その実データに対する挙動を理解する。また、計算手法としての性質を分析し、精度向上のための改良や拡張、いくつかの実グラフデータに対する予測性能や振る舞いの傾向の差を理解する。特に、実験系や用いた化合物系に関する不均質さの影響を理解し、対処法を検討する。

３．研究の方法

(1) 多数のグラフからの教師付き学習として、部分グラフの有無を特徴量とする手法はかなり多くの既存アプローチを包含する設定であるため、この枠組みで考えられる部分グラフ特徴をどのように選択するか、あるいは対象とする部分グラフ特徴集合をどのように限定するか、を、部分グラフの探索を伴う厳密手法を対象に分析する。

(2) 主として実際にドメイン知識を用いてケモインフォマティクス領域で主たる手法となっているデータからフィンガープリントを生成する手法と、グラフからの汎用的教師付き学習として提案されたグラフカーネル法およびブースティングに基づく学習について性質と仕組みを分析・理解し、新たな手法や拡張について検討する。

(3) 対象問題は、データ中に生起する部分グラフを全列挙・全探索する技術と密接な関連を持つため、グラフ列挙やそれに必要となるグラフ同型・部分グラフ同型のアルゴリズムの調査と効率を理解する。特に実データで効率が期待されるヒューリスティクスについて検討を行う。

(4) 不均質性が混合モデリング可能な場合の対処として、多数のグラフデータに対してクラスタリング機構を伴う教師付き学習やデータ分割による教師付き学習のアプローチを検討する。決定論的に教師付き学習を行う場合、不均質であったとしても各々の領域では決定論的判断・予測を行うこととなるため、基本的には特徴空間の各領域ごとでモデルが異なるような手法が必要と考えられる。また、領域ごとに有効な特徴が違いうるため、例えば Supervised LDA のような属性クラスタリングを伴う手法や、組成モデリングによって変数選択を行う手法、決定木・回帰木やランダムフォレストのように領域分割および特徴を伴う手法など、幅広い選択枝を検討する。

(5) PubChem や ChEMBL 等から利用できる様々なアッセイデータをベンチマークデータとして実際に数値実験ができるような系を確立し、分子グラフ表現の選択、データ数・クラスバランス、各種法の実際の予測精度とデータセットの間の傾向、ハイパーパラメタの影響、学習にかかる時間、などの実際的因子について定量的に計測・分析を行う。

４．研究成果

主な研究成果として下記の４点が得られた。

(1) 疎性モデリングに基づく線形学習の一般化と相関構造・冗長性の影響

有力な既存手法として全ての可能な部分グラフ特徴を探索範囲とするが、そのうち与えられた教師付き学習に必要となる部分グラフのみを効率的に同定しながら予測モデルの学習も行うアプローチが提案されている(Kudo et al, 2004; Saigo et al, 2009)。これらは全ての部分グラフの中から最も現在予測モデルの精度を向上させる特徴を加える、という処理を逐次的に行うアプローチであり、この逐次的な変数探索/変数選択はブースティングの枠組みで実現されていた。学習モデルは各ステップで同定される弱学習器である特定の部分グラフの指示子に関して加法的に構成され、指示子の線形モデルとなる。

この枠組みにおいて、どのような教師付

き学習問題が全ての部分グラフ指示子上で求解可能かを一般的な微分可能な損失関数の最小化に 1-norm と 2-norm に関する正則化をつけた形で定式化・分析し、実際に収束が安定的で反復数の少ない学習アルゴリズムを示し論文として発表した(IEEE TPAMI, 2017)。この例として L1 罰則付きのロジスティック回帰モデルを全ての部分グラフ指示子を変数として学習するケースの実際的分析もあわせて行った。

(2) データの「不均質さ」と部分グラフ指示子の特徴空間の特性と非線形性

(1)では従来手法も含めて、部分グラフの有無を表す 0/1 変数に対して線形モデルの学習を対象としていた。しかし、実際の定量的構造活性相関での分子グラフデータでは、データや対象の生物活性に起因する「不均質性」から、このような線形モデルによる表現では限界があるという知見が得られた。実際に、ブースティングにより全部分グラフ指示子で線形学習した場合の精度は ECFP 法などのデータ駆動型の特定の部分グラフ指示子の有無特徴に対して、ランダムフォレスト法や勾配ブースティング法を適用した場合より劣ることもいくつかの実データを通した比較検証から判明した。これにはまず、有無が 0/1 という二値パターンである(Boolean Cube の端点のみに値を取りうる)ことがあげられる。任意の 0/1 変数上の実関数は一意な多項式展開を持つことより、線形モデルで表現可能な仮説クラスはかなり限定的であることも分かる。従って、二次以上の高次の相互作用を含む予測モデリングについて、主に部分グラフ間の共起を列挙する手法および多項式回帰モデルを構築することで分析・検討を行った。引き続き、以降の結果と併せて精密に検討中である。

また、部分グラフの有無という変量は部分グラフ同型性に由来する非常に特殊で強い相関を持つ。特徴空間の非線形性はこの相関構造にも起因すると考えられ、特徴の選択・分解・クラスタリングは重要であると考えられる。選択の観点からは(1)および(3)の観点から分析したが、Supervised LDA 等のクラスタリングを伴う学習モデルとしても引き続き検討中である。

(3) 全部分グラフ指示子に基づく決定木・回帰木の学習とその勾配ブースティング

(2)で得られた特徴空間自体に関する構造的な知見の一方で「不均質性」そのものはそもそも異なる構造活性相関を持つ対象が同一データとして観測されることにも起因すると考えられる。この場合、多種類の傾向の異なる法則を同定する必要があり、混合モデルや領域ごとに異なる法則が学習される再帰的領域分割に基づく手法が有効と考えられる。そこで、全部分グラフ指示子上で決定木・回帰木を学習する厳密アルゴリズムを構築し、実際のデータでその精度評価を行った。また、決定木・回帰木は高次の交互作用や領域ごとのルール生成には有効であるがモデル構造が単純なためアンダーフィットしやすことから、この決定木・回帰木を弱学習器としてブースティングによるアンサンブル学習も併せて構築した。実際にデータセットにもよるが線形手法より精度向上が見られており、引き続き正則化などの細部の検討を行っている。

(4) 部分グラフ指示子のワイルドカードマッチングによる表現緩和

特徴量間の強い相関構造や冗長性および特徴空間の非線形性を緩和吸収する別の試みとして、部分グラフパターンの表現・データグラフとのマッチングの際に、ワイルドカードラベルを許容する拡張について学習アルゴリズムの構築と実際の機械学習における効果の評価を行った。

部分グラフ特徴の有無の特徴ベクトル表現の問題として、部分グラフ特徴間の相対的位置関係が失われることが挙げられるが、ワイルドカードラベルを導入することで複数の部分グラフ特徴が間に任意のラベルの構造を通して同時に起こる状況等をモデル化できると考える。実際にワイルドカード緩和をともなわない場合と比べて線形モデルでは予測精度の向上が見られた。ただし、非線形モデルで学習する場合にはワイルドカードの効果は吸収できる場合が多いことも示唆されており、部分グラフ特徴の厳密な部分グラフ同型を緩和する、という方向性については、引き続き、効率的な方法も検討中である。

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　9　件）

(1) Takigawa I, Mamitsuka H, Generalized sparse learning of linear models over the complete subgraph feature set. IEEE

Transactions on Pattern Analysis and Machine Intelligence. 2017; 39(3): 617-624.

(2) Backhus J, Takigawa I, Imai H, Kudo M, Sugimoto M, An online self-constructive normalized Gaussian network with localized forgetting. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. 2017; E100.A (3): 865-876.

(3) Takigawa I, Shimizu K, Tsuda K, Takakusagi S, Machine-learning prediction of d-band center for metals and bimetals. RSC Advances. 2016; 6: 52587-52595.

(4) 横山侑政・瀧川一学，全部分グラフ指示子に基づく決定木学習．人工知能学会研究会資料 B502，75-80，2016.

(5) Kojaku S, Takigawa I, Kudo M, Imai H, Dense core model for cohesive subgraph discovery. Social Networks. 2016; 44: 143-152.

(6) Nakamura A, Takigawa I, Tosaka H, Kudo M, Mamitsuka H, Mining approximate patterns with frequent locally optimal occurrences. Discrete Applied Mathematics. 2016; 200:123-152.

(7) 瀧川一学，多数のグラフからの統計的機械学習．深化する機械学習：技術の進展とその応用特集号，システム/制御/情報，Vol 60，No 3，2016.

(8) 岡崎文哉・瀧川一学，Wildcard を許容した頻出部分グラフマイニング．電子情報通信学会技術研究報告 115(323)，25-32，2015.

(9) 瀧川一学，データマイニングとしての多重標的相互作用解析．日本薬学会 構造活性相関部会・ニュースレター SAR NEWS No.29，2015.

〔学会発表〕（計 17 件）

(1) Backhus J, Takigawa I, Imai H, Kudo M, Sugimoto M, Reducing Redundancy with Unit Merging for Self-constructive Normalized Gaussian Networks. The 25th International Conference on Artificial Neural Networks (ICANN 2016), Barcelona Spain, September 6-9, 2016.

(2) Backhus J, Takigawa I, Imai H, Kudo M, Sugimoto M, Online EM for the Normalized Gaussian Network with Weight-Time-Dependent Updates. The 23rd International Conference on Neural Information Processing (ICONIP 2016) Kyoto, Japan, October 16-21, 2016.

(3) 瀧川一学，科学と機械学習のあいだ:変量の設計・変換・選択・交互作用・線形性．第 19 回情報論的学習理論ワークショップ(IBIS2016)，京都大学，平成 28 年 11 月 16 日-19 日.

(4) 岡崎文哉・瀧川一学，Wildcard 許容特徴量のグラフ分類における効果の分析．第 19 回情報論的学習理論ワークショップ(IBIS2016)，京都大学，平成 28 年 11 月 16 日-19 日.

(5) 穐本浩昇・田中 譲・瀧川一学，ABS 作動データを用いた分析による札幌市内の道路凍結の予測.情報処理学会第 79 回全国大会，3V-07，名古屋大学，平成 29 年 3 月 16 日-18 日.

(6) 鈴木慶介・今井英幸・張 若霓・瀧川一学・湊 真一，平行移動不変な非負値行列因子分解とその分析．第 15 回情報科学技術フォーラム(FIT2016)，富山大学，平成 28 年 9 月 7 日-9 日.

(7) 岡崎文哉・瀧川一学，Wildcard 許容頻出部分グラフパターンのグラフ分類への応用．2016 年度人工知能学会全国大会(JSAI2016)，3E4-3，北九州国際会議場，平成 28 年 6 月 6 日-9 日.

(8) Jana Backhus・瀧川一学・今井英幸・工藤 峰 一 ・ 杉 本 雅 則 ， An Online Self-constructive Locally Updated Normalized Gaussian Network with Localized Splitting. 2016 年度人工知能学会全国大会(JSAI2016)，3E4-3，北九州国際会議場，平成 28 年 6 月 6 日-9 日.

(9) 瀧川一学，(招待講演) データマイニングとしての多重標的相互作用解析．CBI 学会 2015 年大会，FS-08，in silico によるポリファーマコロジー創薬，タワーホール船堀，平成 27 年 10 月 28 日.

(10) 瀧川一学，(招待講演) データマイニングとしての多重標的相互作用解析．第 365 回 CBI 学会講演会，フェノタイプスクリーニング 古くて新しい創薬手法 Part2，東京工業大学キャンパス・イノベーションセンター，平成 27 年 7 月 9

日.

(11) 横山侑政・瀧川一学, 全部分グラフ指示子に基づく決定木学習. 人工知能学会第 99 回人工知能基本問題研究会(SIG-FPAI), 湯の原ホテル, 平成 28 年 1 月 21 日-22 日.

(12) 岡崎文哉・瀧川一学, Wildcard を許容した頻出部分グラフマイニング. 第 18 回情報論的学習理論ワークショップ(IBIS2015), つくば国際会議場, 平成 27 年 11 月 25 日-28 日.

(13) Tanaka A, Takigawa I, Imai H, Kudo M, Theoretical analyses on ensemble and multiple kernel regressors. The 6th Asian Conference on Machine Learning (ACML2014), Nha Trang, Vietnam, November 26-28, 2014

(14) Tanaka A, Takigawa I, Imai H, Kudo M. Analyses on generalization error of ensemble kernel regressors. Proceedings of the Joint IAPR International Workshop on Statistical, Structural, and Syntactic Pattern Recognition (S+SSPR 2014), Joensuu, Finland, August 20-22, 2014.

(15) 瀧川一学, 疎性モデリングに基づく部分グラフ特徴学習. ERATO 湊離散構造処理系プロジェクト 2014 年度 秋のワークショップ, 北海道礼文島ピスカ 21, 平成 26 年 9 月 7 日～9 月 10 日.

(16) 瀧川一学, 疎性モデリングに基づく部分グラフ特徴学習. 第 17 回情報論的学習理論ワークショップ(IBIS2014), 名古屋大学・名古屋工業大学, 平成 26 年 11 月 16 日～11 月 19 日.

(17) 瀧川一学, (招待講演) 多数のグラフからの統計的機械学習. 人工知能学会 第 94 回 人工知能基本問題研究会(SIG-FPAI), 根室市総合文化会館, 平成 26 年 7 月 24 日.

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況(計 0 件)

○取得状況(計 0 件)

６．研究組織

(1)研究代表者

瀧川 一学 (TAKIGAWA, Ichigaku)
北海道大学大学院情報科学研究科・准教授
研究者番号：10374597