科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２８　日現在

機関番号：　１４３０１
研究種目：　研究活動スタート支援
研究期間：　2012～2013
課題番号：　２４８０００３６

研究課題名（和文）機械学習アルゴリズムのための最適グラフ構成法に関する研究

研究課題名（英文）A study on optimizing graphs for machine learning algorithms

研究代表者
烏山　昌幸（KARASUYAMA, Masayuki）

京都大学・化学研究所・助教

研究者番号：４０６２８６４０

交付決定額（研究期間全体）：（直接経費）　　2,300,000円、（間接経費）　　690,000円

研究成果の概要（和文）：生物学におけるタンパク質の相互作用ネットワークやソーシャルネットワークにおけるリンク構造など，様々な場面でネットワークとして表現されるデータの解析に注目が高まっている．このようなネットワークは数学的には「グラフ」と呼ばれる．本研究ではグラフとしての特徴をもつデータを扱う基本アルゴリズムの開発を行ったものであり，グラフ上での予測問題において既存の手法とくらべ高い精度を持つことを実験的に確認することに成功した．

研究成果の概要（英文）：A variety types of network data have been attracted wide attention, such as protein interaction network in biology and link relationships in social networks. This research has studied statistical algorithms to analyze these network data, which can be represented as ``graph'', and developed highly accurate methods for prediction problem on graphs compared to existing approaches.

研究分野：　総合領域

科研費の分科・細目：　情報学・知能情報学

キーワード：　機械学習　グラフ　バイオインフォマティクス

１．研究開始当初の背景

ハードウェア技術の急速な発展に伴い，様々な分野において大量のデータをコンピュータ上で管理することが可能になってきている．一方で蓄積したデータから如何に有用な情報を抽出するかが重要な課題となり，機械学習あるいはデータマイニングと呼ばれるデータの背後に潜む規則をコンピュータによって自動的に発見する技術の注目度が高まっている．

近年では，WWW ネットワーク上のドキュメントやタンパク質の相互作用データなどの個人や物質間の関係性を表現した構造データの解析が重要性を増している．このようなデータは単純な数値データを対象とする古典的な統計的手法では扱い辛い場合があり，方法論レベルでの研究が盛んに行われている．

関係性を表現する方法として「グラフ」に基づく方法がその汎用性の高さから広く利用されている．一旦データをグラフとして表現してしまえば，グラフ上での推定問題として様々な問題を定式化できる．たとえば，バイオインフォマティクス分野ではタンパク質の機能予測という問題が知られている．相互作用のネットワーク上で隣接しているタンパク質は似通った機能を持つ事が多いと言われている．そのためネットワーク内で機能既知のタンパク質から機能未知のタンパク質の機能をネットワークの接続構造に沿って予測する問題として定式化することができる．

このようにグラフ上での推定問題は有用性の高い基本問題として知られている．ところが，決定的な解決策のない実用上の問題もいくつか知られている．たとえば，入力として複数のグラフが与えられた場合の推定や，事前に設定するパラメータの存在などである．本研究はこれらの問題に対して機械学習アルゴリズムに対する最適なグラフの構築という観点から取り組んだものである．

２．研究の目的

グラフ上での推定アルゴリズムでは，どのようなグラフが与えられるかが最終的な結果の善し悪しを大きく左右することになる．本研究では，特にグラフ上の予測アルゴリズムのための最適グラフの構成問題として(1)複数グラフからの最適グラフ構成法と，(2)特徴表現からの最適グラフ構成法，の２つの問題に着目して研究を行った．どちらもグラフ型のデータ解析において頻出する実用上重要な問題であり，本研究では精度の高く，利用の簡単な方法論の確立を目指した．

３．研究の方法

本研究は基礎方法論に関する研究であり，各課題において方法論の設計に重点を置き研究を行った．特に，グラフ上での推定手法として広く使われている Label Propagation アルゴリズム(図 1)を土台にした枠組みを考えることで汎用性の高い手法の構築を目指した．開発した手法に関しては順次計算機実験によって精度を検証し，それぞれの手法の挙動を観察し考察した．また成果に関しては分野内の主要な国際学会，国際学術雑誌において発表を行った．

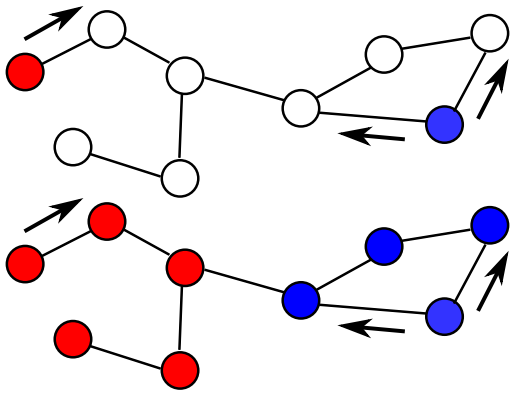

図１：グラフ上の予測問題の概念図（Label Propagation）．色の付いたノードの情報(既知の情報)を伝播させ，未知のノードの情報を推定する．色の違いは例えばタンパク質の機能の違いに相当する．

４．研究成果

主な成果について(1)複数グラフからの最適グラフ構成法と(2)特徴表現からの最適グラフ構成法のそれぞれについて述べ，最後に全体のまとめとする．

(1) 複数グラフの最適結合法

データの収集が容易になった一方で１つのタスクについて複数種類のデータが取得できるケースが増加し，複数あるグラフの中でどの組み合わせが予測に対して真に有用であるのかは事前には分からない場合がある．例えば，タンパク質のデータに対しては相互作用のグラフや，遺伝子発現量，配列データから抽出した情報が有用であると考えられるが，データの計測ノイズの大小など予測タスクについてどのグラフがどの程度有効であるかは自明ではない場合がほとんどである．あるいは発現量や配列データなどの数値，文字列データから類似度情報を基にグラフ構造を考える場合には，どのような類似度尺

度を用いるかでもグラフの構造は変化する．

本項目では，与えられた複数のグラフから予測に必要なものを選び出して最適に結合するための枠組みを提案した．提案法の特徴は与えられたグラフ群のなかには予測にとって有益でないものが含まれている可能性を考慮し，そのようなグラフを除去することにある．どのグラフが予測に有効か事前にわからない状況において，利用できる可能性がある情報を全てアルゴリズムに入力したのであれば実際には必要がないグラフが含まれている可能性が高く，このような特徴によって真に関連の深いグラフを見つけることは予測性と解釈性の両観点から重要となる．

提案法ではスパース正則化と呼ばれる機械学習の手法を応用して，不要なグラフを除去する枠組みを定式化した．具体的にはまず全てのグラフに対して重み付き平均をとり，結合されたグラフでの推定を考える．その後，スパース正則化の枠組みに従って重みを推定すると予測への寄与が小さいグラフの重みが0になり，そのグラフは消去される．この性質について数学的な考察も行い，グラフがどの程度似ていれば似た重みが得られるかを評価する理論などを導いた．複数のグラフから予測を行う手法はすでにいくつか提案されていたが，このように有害なグラフの存在を陽に考慮したものは少なく，シミュレーション実験によってこのような状況においては予測精度が低下する場合があることを確認した．また，遺伝子データやWebテキストデータを用いた精度比較も行い提案法の有効性を検証した．

この成果についてはアメリカの電気電子工学系で最大の学会である IEEE の国際ジャーナル IEEE Transactions on Neural Networks and Learning Systems に掲載された．

(2) 特徴表現からの最適グラフ構成法

グラフに基づく機械学習の予測アルゴリズムはグラフデータのみならず，古典的な数値データからグラフを類似度に基づいて生成して利用することもでき，例えば画像処理などの分野でこのようなアプローチが有効に働くことが知られている．しかし，この場合数値データからグラフを生成する方法には未だ確立された決定的な方法はなくヒューリステックな方法が用いられることが多い．ここでは数値データから生成されたグラフ上での予測問題を考えたときに，どのような方法でグラフを設計するべきなのかという問題に取り組んだ．

我々はグラフが数値データに潜在的に潜むマニフォルドとよばれる低次元構造を近似しているものとして解釈が出来ることに着目した．元の数値データからはこのような構造を見つけることは自明ではないが，数値データの空間内で例えば近傍にある点同士をつないだグラフはそのような潜在的な関係性を明らかにすることが知られている．さらにラベル推定の観点からグラフが元の空間内での類似度を反映するようにした上でグラフを最適化するモデルを定式化した．このアプローチでは事前に設定が必要なパラメータの数が従来の方法より少なく，最適化されたグラフが多くのデータで精度の高い予測を可能にすることを実験的に検証した．

この成果については国際学会 Neural Information Processing Systems で発表を行った．この学会は機械学習の基礎方法論に関するものとしては最高水準のものとして知られている．また現在，実験と解析を追加した内容を学術雑誌に投稿を行っている．

(3)まとめ

本研究ではグラフに基づく機械学習における，主に予測問題に関して，最適なグラフの構成アルゴリズムについて考えてきた．複数グラフの結合と，特徴表現からのグラフ構成に関しては上記の通りある程度成果をまとめることができた．残る重要な課題としてこれらの枠組みを一つに統合するということが挙げられる．つまり，複数のグラフ，複数の数値データが与えられたときどのようにそれらを一括に取り扱い，グラフ上での予測アルゴリズムに渡すのが良いかという問題である．こういった課題をさらにすすめ，バイオインフォマティクスなどの実問題への適用が今後の課題である．

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　１件）

1 M. Karasuyama and H. Mamitsuka, Multiple Graph Label Propagation by Sparse Integration, IEEE Transactions on Neural Networks and Learning Systems, vol. 24, no.12, pp. 1999-2012, 2013. doi: 10.1109/TNNLS.2013.2271327［査読有］

〔学会発表〕（計　３　件）

1 M. Karasuyama and H. Mamitsuka, Manifold-based Similarity Adaptation for Label Propagation, *Advances in Neural Information Processing Systems (NIPS) 26*, pp.1547-1555, 2013.［査読有］

2 烏山昌幸，馬見塚拓，局所線形近似に基づくラベル伝播のための類似度適合，情報論的学習理論と機械学習研究会(IBISML),

信学技報, vol. 112, no. 454, pp.115--121, 2013. [査読無]

3 烏山昌幸, 馬見塚拓, ラベル伝播アルゴリズムにおける複数グラフのスパース結合法, 情報論的学習理論と機械学習研究会(IBISML), 信学技報, vol. 112, no. 279, pp. 171--178, 2012.[査読無]

〔図書〕(計　0　件)

〔産業財産権〕
○出願状況(計　0　件)

○取得状況(計　0　件)

〔その他〕
特になし

６．研究組織
(1)研究代表者
烏山　昌幸　(KARASUYAMA, Masayuki)
京都大学・化学研究所・助教
研究者番号：40628640