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　　５　月　３１　日現在

機関番号：　１３９０３
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３０１２０

研究課題名（和文）グラフ構造データからの部分構造抽出法の開発

研究課題名（英文）A machine learning based approach to analysing latent substructure of graph data

研究代表者

烏山　昌幸（Karasuyama, Masayuki）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：４０６２８６４０

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：多様化するデータの中で、関係性をグラフとして表現できるものは多い。そのようなデータを解析する場合に、グラフ上のパラメータをどのように設定するか、グラフから有用な構造をどのように抽出するかなどが実用上の問題となる。本研究では機械学習におけるグラフに基づく手法の研究を行った。とくに、グラフ上でのラベル推定問題に関するグラフ重みの最適化やグラフからの部分構造の発見に関する手法開発を行った。ラベル推定問題は生物データ解析など応用が多く、精度の高い手法の構築が重要とされる。また、グラフの部分構造発見については組み合わせ的な性質から計算が難しいが、計算効率の高い手法の開発に取り組んだ。

研究成果の概要（英文）：A variety of data can be represented as a graph. For statistical data analysis based on graphs, it is important to consider setting appropriate parameters of graphs and extracting informative structure. This study focuses on a methodological study of ``machine learning'' based on graph data. In particular, the label estimation problem on a graph and the important subgraph identification problem have been considered. For example, accurate label estimation methods and scalable subgraph identification methods are required by biological data analysis.

研究分野：　機械学習

キーワード：　機械学習　グラフ　多様体学習　半教師学習

１．研究開始当初の背景

ビッグデータという言葉が広く浸透し，多様化するデータをどう活用するかに強い社会的な関心が集まっている．そこで，データから背後に潜む規則性を推定する機械学習やデータマイングと呼ばれる技術の重要性が高まっている．

本研究では特に，関係性を持つデータの表現方法としてのグラフに着目する．グラフを使うことで個々の対象をノード，対象間の関係をエッジとして表現することができる．例えば，生物学分野では個々のタンパク質をノードとして表し，相互作用をエッジで表現するタンパク質相互作用ネットワークを考えることで生体内のメカニズムを解析することがある．あるいは，web ページは各ページがハイパーリンクで繋がったグラフだと捉えることができる．

このように表現されたグラフデータのうえで，様々なデータ解析の課題を考えることができる．例えば，各ノードにラベルを割り振り，それを予測するラベル推定問題は汎用性が高い．タンパク質のネットワークでは，各タンパク質の機能カテゴリーをラベルだとすると，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予測する問題を考えることができる．あるいは，化合物のようにグラフと表現できるデータが大量にある場合もある．このとき，化合物の性質（毒性など）に対してグラフのどの部分が影響しているのか検出する問題は化学データ解析にしばしば表れる．これらの問題では，古典的な数値テーブルとして表現されるデータが付随することも多く，それらをどのように統合的に扱うかも重要な問題となる．また，グラフの部分構造を考えると，ありうる部分構造全てを考慮すると組み合わせ的な計算困難性が表れることもある．

２．研究の目的

上で述べたように，多様化するビッグデータに対して，グラフは関係性を表現する汎用ツールとなる．一方で， 数値データとの統合や，部分構造の抽出などは未だ決定的な解決策はなくヒューリスティックなアプローチが頻繁に用いられる．本研究では，グラフに基づくデータの表現と，機械学習における多様体学習やセーフスクリーニングと呼ばれる技術を組み合わせて，高精度な予測手法や高効率な部分構造発見手法の開発を行った．これらは，グラフに基づくデータ解析の新たな枠組みの構築を目指したものである．

３．研究の方法

本研究は方法論研究であり，いくつかの基本問題設定に手法を数理的に定義，検証し，ベンチマークデータなどによる精度実証を行った．具体的には，グラフ上のラベル推定問題や重要部分グラフ発見問題，グラフによる数値データの潜在構造解析などに対する基礎方法論の確立に取り組んだ．成果については，分野内の主要な国際論文誌，国際会議などで精力的に発表を行った．

４．研究成果

はじめに，グラフ上のラベル推定問題について述べる．グラフ上のラベル推定問題ではエッジの接続関係を手掛かりに，グラフのノードに部分的に付けられたラベルからラベルのわからないノードのラベルを予測する（機械学習では半教師付き学習と呼ばれる問題に分類される）．本研究では特に数値データから関係性を表現するグラフを構築する場合を考える．この場合，グラフは数値データに潜む多様体構造を近似するものとして解釈できることが知られている．ここでの多様体構造とは高次元の空間中に存在する低次元の潜在的なデータの分布のことを指す．高次元のデータは一般にパラメータ推定が難しいことが知られているが，実際には本質的には低次元の自由度しか持たないデータは多く，このようなアプローチを考えることがある（例えば，ある数字の手書き数字のピクセルデータはピクセル数の次元と比べて本質的な自由度はずっと低い場合が多い）．この解釈に基づき，グラフのエッジに割り振られた類似性の尺度を多様体の構造を良く近似するように最適化することで，ラベル推定の精度が向上することを示した．また，提案する枠組みが，ラベル推定問題だけでなくクラスタリングにも適用可能なことを示した．以上の内容について，画像データなどのベンチマークで精度を検証し，機械学習の主要な国際雑誌である，Machine Learning 誌で発表した．

次に，グラフから特定の性質を予測するモデルを構築し，予測に重要な寄与を持つ部分構造を発見するための方法について述べる．この問題はありうる部分グラフの数の爆発により，計算が困難になることが多い．ここでは，機械学習において予測モデルから重要要素を発見する場合に用いられる LASSO と呼ばれる手法（スパース学習）をベースに検討をした．この方法はパラメータの大部分を 0 にすることで効率的な計算が可能だが，部分グラフ発見の設定では，この手法でも最適化計算が困難になる．本研究では，LASSO において，不要な部分グラフを効率的かつ近似なしに正確に削除する方法を機械学習のセーフスクリーニングをベースに構築した．提案法が化合物の部分構造発見問題などについて，効率的な計算を可能にすることを示し，データマイニングの主要な国際会議である KDD で

発表を行った.

多様体学習に関する研究では異なる変数間の依存関係を解析する方法論の検討を行った．変数の依存関係の解析は古くから研究のあるトピックであるが，複雑な非線形の関係性を考える場合には未だ計算の難しい問題である．ここでは，依存関係のある変数は共通の低次元の多様体構造を共有するという仮定のもと変数のクラスターを求める手法を提案した．多様体構造は数値データを近似するグラフとして推定され，共通の潜在構造を持つ変数の集合は共通のグラフによって近似される．現在，プログラムの構築とベンチマーク実験を行っている．部分的な成果について国内学会などで発表を行っている．

以上に加えて，グラフ表現関連手法の普及を目指して，学会での講演を行った．また，機械学習の基礎技術普及を目指して，書籍の翻訳も行った.

以上が，本研究で取り組んだ主要な課題のこれまでの結果である．ラベル推定問題，部分構造発見法に関しては査読付きの主要な関連雑誌，学会での発表に至った．多様体学習に関しては現在も進行中であり，アルゴリズムのスケーラビリティの改善に取り組んでいる．科研費としての期間は終了するが今後，学会，雑誌での発表を目指して継続的に進めていく．

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）
M. Karasuyama, and H. Mamitsuka, Adaptive Edge Weighting for Graph-Based Learning Algorithms, *Machine Learning*, vol.106, issue 2, 307-335［査読有］

〔学会発表〕（計　7件）
K. Nakagawa, S. Suzumura, M. Karasuyama, K. Tsuda, and I. Takeuchi, Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1785--1794, San Francisco, USA, Aug. 13-17, 2016.［査読有］

烏山昌幸，多様体学習に基づく特徴クラスタリング，応用数理学会 2016 年会，オーガナイズドセッション「機械学習」， 2016 年 9 月 12 日

烏山昌幸，機械学習による粒界データ解析，応用数理学会 2016 年会，オーガナイズドセッション「マテリアルズインフォマティクスと応用数理」， 2016 年 9 月 14 日

烏山昌幸，馬見塚拓， 複数多様体の同時推定に基づく特徴クラスタリング， 情報論的学習理論と機械学習研究会(IBISML)， 信学技報，vol. 115, no. 323, IBISML2015-79, pp. 195-201, 2015 年 11 月

烏山昌幸，機械学習によるグラフデータ解析，日本生体医工学会大会 企画シンポジウム「神経科学と信号処理の邂逅」， 2015 年 5 月 9 日

M. Karasuyama, and H. Mamitsuka, Learning Kernel-based Feature Representation for Gene Essentiality Prediction, *RECOMB/ISCB Conference on Regulatory and Genomics with DREAM Challenges and Cytoscape Workshops* 2014， 2014 年 11 月 11 日

烏山昌幸，馬見塚拓，Manifold-based Similarity Adaptation for Label Propagation, 第 17 回 画像の認識・理解シンポジウム(MIRU2014)， 2014 年 7 月 29 日

〔図書〕（計　１ 件）

統計的学習の基礎 -データマイニング・推論・予測- （１２章担当），共立出版，2014.
原著：Trevor Hastie, Robert Tibshirani, Jerome Friedman, 監訳：杉山 将，井手 剛，神嶌 敏弘，栗田 多喜夫，前田 英作，翻訳：井尻 善久，井手 剛，岩田 具治，金森 敬文，兼村 厚範，烏山 昌幸，河原 吉伸，木村 昭悟，小西 嘉典，酒井 智弥，鈴木 大慈，竹内 一郎，玉木 徹，出口 大輔，冨岡 亮太，波部 斉，前田 新一，持橋 大地，山田 誠

〔産業財産権〕

○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
烏山昌幸（KARASUYAMA, Masayuki）
名古屋工業大学・情報工学科・准教授
研究者番号：40628640