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 ２７ 年　 ９ 月 ３０ 日現在

機関番号：１４３０１
研究種目：若手研究(A)
研究期間：2010～2014
課題番号：２２６８００１２

研究課題名（和文）グラフ構造データの予測的分析のための機械学習手法の研究

研究課題名（英文）Predictive Machine Learning Methods for Graph-structured Data

研究代表者
鹿島　久嗣（Kashima, Hisashi）

京都大学・情報学研究科・教授

研究者番号：８０５４５５８３

交付決定額（研究期間全体）：（直接経費）　　11,400,000円

研究成果の概要（和文）：機械学習をはじめとするデータ分析技術において、その従来手法の多くは、分子構造や特許文書、交友関係や企業間取引ネットワークなど、データの内や外にグラフ構造をもつような対象を扱うことが苦手である。
本研究では、意思決定において重要な役割を果たす予測モデリングの観点から、データの内外に存在する様々に複雑なグラフ構造データを統合的・効率的・効果的に扱うための種々の方法論を開発した。

研究成果の概要（英文）：Existing data analysis methods including machine learning are not readily designed for complex data with inside/outside graph structures such as chemical compounds, patent documents, social networks, and business networks.
In this this research project, we developed integrated, efficient, and effective methods for such complex graph-structured data from the predictive modeling perspective, which play key roles in decision making.

研究分野：機械学習

キーワード：機械学習　人工知能　データマイニング　グラフ構造データ　ネットワークデータ　関係データ　予測

1．研究開始当初の背景

データに潜む規則性を発見し、これを新たな知見の創出や意思決定に役立てていくためのデータ分析技術の研究は、機械学習やデータマイニング、パターン認識、統計科学など様々な分野で精力的に進められてきた。しかし、近年これらの枠では捉えられない複雑な構造をもったデータが、多くの重要な場面において急速に増加している。多くの場合、それらは最も高い記述力をもった形式のひとつであるグラフ構造によって記述される。例えば WWW は HTML 文書等が互いにハイパーリンクと呼ばれる関係で繋がれたものであり、グラフ構造を持っている。また、各々の HTML 文書の中には、やはりタグの包含関係などを示したグラフ構造が存在する。学術文献や特許文書なども同じくグラフ構造を内在する XML 形式によって表現され、文書間の参照関係もまた別のグラフ構造をなしている。遺伝子やタンパク質の配列は、グラフ構造の一種である配列構造として、化合物などの構造などは原子とそれを結ぶ共有結合を表したグラフ構造として表される。さらに、遺伝子やタンパク質と化合物との間の相互作用などの関係もまた、上位のグラフ構造を形成している。 一方、ビジネスの文脈では、近年盛んに利用されているソーシャルネットワークサービス等では人間同士の交流関係を表したグラフ構造が、さらには、彼らが属する企業内の部門の階層構造、企業間の取引関係など、複数階層にまたがる多種多様のグラフ構造が形成されている。オンラインショッピングサイトなどでは、顧客に適切な商品を推薦する仕組みが成功を収めているが、これらもまた、人と人、顧客と商品の間の関係を表現したグラフ構造として表される。

このように、グラフ構造をもったデータは、Web マイニング、バイオ／創薬、ビジネス分析、マーケティングなど、実世界の多くの重要な場面において自然に現れてくるが、現状では、これらを統一的に扱うことのできる分析手法はまだ十分に整備されているとは言い難い。たとえば、従来提案されている殆どの分析手法では、扱うことのできるデータの形式はベクトル形式、すなわち一定数の決まった項目が並んでいる形式が仮定されており、グラフ構造を直接扱うことは想定されていない。 また、グラフ構造には化合物、Web ページ、特許文書、企業などの注目するデータ単位に内在するグラフ構造と、それらの間の繋がりを表す外側のグラフ構造の 2 種類があるが、これまでの研究は、注目するデータ単位の内と外いずれかの構造にのみ注目して行われてきており、これらを統一的に扱うことのできる方法は少ない。さらに、グラフという離散的な構造であるということに起因する組み合わせ的な性質が、計算量の爆発を引き起こす。

データ分析手法の種別を考えた場合、「いま何が起きているのか」を理解するための現状把握的データ分析と「これから何が起こるのか」を予測する予測的データ分析の 2 種類があるが、グラフマイニングやネットワーク分析などの分野が主に対象としていたのは前者であるのに対し、より直接的に競争力のある意思決定に結び付くのは後者の予測的データ分析である。

以上の課題は機械学習やデータマイニングのコミュニティにおいても近年議論が盛んである。

2．研究の目的

本研究では、内外にグラフ構造を有するデータの予測的モデリング方法論の確立を目標とし、以下の 4 項目に取り組む：

(1) 理論：
グラフ構造の予測モデリングを対象とした学習理論

(2) アルゴリズム：
数十万～数百万のノードをもつグラフに対し現実的な時間で予測を行うことのできる対規模性の高い効率的なアルゴリズム

(3) 応用：
各分野の専門家と協力した、各ドメインにおけるグラフ構造解析への取り組み

(4) 情報発信：
上記研究成果についての情報発信

グラフ構造データを対象とした予測的な分析における基本問題としては以下の２つがあり、両者に取り組む：

(5) ノード予測問題：
化合物、特許文書、企業などの注目するデータ単位のもつ性質を予測する問題

(6) リンク予測問題：
化合物、Web ページ、企業などの注目するデータ単位の間の関係（リンク）を予測する問題

いずれの課題も各方面における専門家と協力して進める。

3．研究の方法

理論・アルゴリズム開発においては、グラフカーネル法や行列分解に基づくアプローチなど、これまでに我々が世界に先駆けて研究を行ってきたグラフ構造データを対象とした解析方法を基礎に、その性能の理論的な解析ならびに耐規模性の高いアルゴリズムの開発に取り組む。その際、学習理論や行列・グラフアルゴリズムの専門家とともにこれらに取り組むことで研究を加速する。

応用においては各分野の専門家の協力のもと現実的なインパクトの高い問題を特定

し、各ドメインの情報を定式化やアルゴリズムの設計にうまく取り込みつつ取り組む。
また、国内外の最新の関連研究動向の情報収集を積極的に行ない、プロジェクトの情報発信を適切な時期に有効に行なえるよう注意を払う。
研究の実施では、基礎理論構築やアルゴリズム開発の段階で平行して行なえる部分を見出し、研究プロジェクト全体の進行が遅延しないよう特に気をつける。たとえばノード予測問題とリンク予測問題の定式化とアルゴリズム設計も比較的独立して行えるため、それぞれの進み具合に応じて、どちらかを先に応用に取り組むなども可能であり、計画の遂行状況に応じて柔軟に対応する。また、研究の初期から積極的に研究パートナーと連携をとり、早い段階からのデータ準備や、基礎理論に基づいたアルゴリズムを早急に適用できるように準備を整えることを心がける。 また、それにより実応用における知見を基礎理論に取り込むことも期待する。必ずしも、理論→アルゴリズム→応用の順に完成させるというものではなく、応用→アルゴリズム→理論というフィードバックも取り入れつつ柔軟に対応する。

4．研究成果

研究の目的で述べた4項目それぞれに対する研究成果は以下の通りである：

(1) 理論：
グラフ構造をもったデータに対する学習理論にあたり本質となるのは、グラフの辺にあたる「データの対」である。我々はデータ対同志の類似度（カーネル関数）を提案し、これを用いた予測精度の理論的解析を行った。
また、グラフ構造データを対象とする新しい学習問題の提案も行い、グラフ構造データ解析の新たな可能性を示した。

(2) アルゴリズム：
グラフの典型的な表現方法のひとつが隣接行列による表現であり、大規模なグラフ構造データを扱うためには行列を対象としたアルゴリズムの開発が必要である。多くの実グラフデータは数学的には低ランク性と呼ばれる性質をもっており、この性質を利用することによって、少ないデータから高精度で、かつ非常に効率よく計算できるアルゴリズムを開発した[5,6]。また、グラフ上での大規模な予測問題を解くために 2003 年に我々が世界で初めて提案して以来、多くの研究者によって発展されてきたグラフカーネル法を、さらに高速化させることで大規模データに対しても適用できるようにした。
また、グラフの重要な部分クラスである木構造をもつデータに対し、その構造の特殊性を利用して、より高速、高精度で解析を行う手法も開発した。同様に配列構造の予測問題に対しても対規模性の高い方法を提案した。

(3) 応用：
主にバイオ・化学分野においてグラフ構造データ解析の応用を行った。その対象はタンパク質の複合体予測や薬剤と標的分子の相互作用予測、化学反応の予測など多岐にわたる。

(4) 情報発信：
グラフ構造データを対象としたデータ解析法についての招待講演やセミナー等を通じた情報発信を行った。また、本研究の成果は学会等での各種賞を受賞するなど、国内外で高い評価を得た。

5．主な発表論文等

〔雑誌論文〕（計４件）

① Hisashi Kashima, Satoshi Oyama, Yoshihiro Yamanishi, Koji Tsuda. Cartesian Kernel: An Efficient Alternative to the Pairwise Kernel. *IEICE Transaction on Information and Systems*, Vol.E93-D, No.10, pp.2672-2679, 2010. http://search.ieice.org/bin/summary.php?id=e93-d_10_2672

② Yosuke Ozawa, Rintaro Saito, Shigeo Fujimori, Hisashi Kashima, Masamichi Ishizaka, Hiroshi Yanagawa, Etsuko Miyamoto-Sato, Masaru Tomita. Protein Complex Prediction via Verifying and Reconstructing the Topology of Domain-domain Interactions. *BMC Bioinformatics*, Vol. 11, No. 350, 2010. doi:10.1186/1471-2105-11-350

③ Reiji Teramoto, Hisashi Kashima. Prediction of Protein-ligand Binding Affinities Using Multiple Instance Learning. *Journal of Molecular Graphics and Modelling*, Vol.29, No.3, pp.492-497, 2010. doi:10.1016/j.jmgm.2010.09.006

④ 木村 大翼，久保山 哲二，渋谷 哲朗，鹿島 久嗣．部分パスに基づいた木カーネル．人工知能学会論文誌，Vol.26，No.3，pp.473-482，2011. doi:10.1527/tjsai.26.473

〔学会発表〕（計９件）

① Rudy Raymond, Hisashi Kashima: Fast and Scalable Algorithms for Semi-supervised Link Prediction on Static and Dynamic Graphs, In *Proc. European Conference on Machine Learning and Principles and*

Practice of Knowledge Discovery in Databases (ECML PKDD), pp.131-147, Barcelona, Spain, 2010.
doi: 10.1007/978-3-642-15939-8_9

② Ryota Tomioka, Taiji Suzuki, Masashi Sugiyama, Hisashi Kashima: A Fast Augmented Lagrangian Algorithm for Learning Low-Rank Matrices, In *Proc. 26th International Conference on Machine Learning (ICML)*, pp.1087-1094, Haifa, Israel, 2010.
http://www.icml2010.org/papers/556.pdf

③ Mutsumi Fukuzaki, Mio Seki, Hisashi Kashima, Jun Sese: Finding Itemset-Sharing Patterns in a Large Itemset-Associated Graph, In *Proc. 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp.147-159, Hyderabad, India, 2010.
doi: 10.1007/978-3-642-13672-6_15

④ Daisuke Kimura, Tetsuji Kuboyama, Tetsuo Shibuya, Hisashi Kashima: A Subpath Kernel for Rooted Unordered Trees, In *Proc. 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp.62-74, Shenzhen, China, 2011.
doi: 10.1007/978-3-642-20841-6_6

⑤ Yuta Tsuboi, Yuya Unno, Hisashi Kashima, Naoaki Okazaki: Fast Newton-CG Method for Batch Learning of Conditional Random Fields, In *Proc. 25th AAAI Conference on Artificial Intelligence (AAAI)*, pp.489-494, San Francisco, California, USA, 2011.
https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewFile/3448/3876

⑥ Daisuke Kimura, Hisashi Kashima: Fast Computation of Subpath Kernel for Trees, In *Proc. 29th International Conference on Machine Learning (ICML)*, pp.393-400, Edinburgh, Scotland, 2012.
http://arxiv.org/ftp/arxiv/papers/1206/1206.4642.pdf

⑦ Shohei Hido, Hisashi Kashima: Hash-based Structural Similarity for Semi-supervised Learning on Attribute Graphs, In *Proc. 23rd International Conference on Pattern Recognition (ICPR)*, pp.3009-3012, Tsukuba, Japan, 2012.
http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6460798

⑧ Yoshifumi Aimoto, Hisashi Kashima: Matrix Factorization with Aggregated Observations, In *Proc. 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp.521-532, Gold Coast, Australia, 2013.
doi: 10.1007/978-3-642-37456-2_44

⑨ Jingjing Wang, Satoshi Oyama, Masahito Kurihara, Hisashi Kashima. Learning an Accurate Entity Resolution Model from Crowdsourced Labels. In *Proc. the 8th International Conference on Ubiquitous Information Management and Communication (ICUIMC/IMCOM)*, Siem Reap, Cambodia, 2014.
doi:10.1145/2557977.2558060

〔図書〕（計３件）

① Yoshihiro Yamanishi, Hisashi Kashima: Prediction of Compound-protein Interactions with Machine Learning Methods, In *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pp.304-317, IGI Global, 2010.
doi:10.4018/978-1-61520-911-8.ch016

② Hisashi Kashima, Hiroto Saigo, Masahiro Hattori and Koji Tsuda: Graph Kernels in Chemoinformatics, In *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pp.1-15, IGI Global, 2010.
doi:10.4018/978-1-61520-911-8.ch001

③ Tetsuo Shibuya, Hisashi Kashima, Jun Sese and Shandar Ahmad (編): *Pattern Recognition in Bioinformatics, Proceedings of the 7th IAPR International Conference (PRIB 2012), Lecture Notes in Computer Science*, Vol.7632, 2012.
http://www.springer.com/us/book/9783642341229

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕
なし

6．研究組織
(1) 研究代表者
鹿島 久嗣 (KASHIMA, Hisashi)
京都大学・大学院情報学研究科・教授
研究者番号：80545583

(2) 研究分担者

なし

(3) 連携研究者
なし

(4) 研究協力者
なし