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　２８　年　１０　月　　５　日現在

機関番号：　１４３０１
研究種目：　基盤研究(B)（一般）
研究期間：　2012～2015
課題番号：　２４３０００５４

研究課題名（和文）多様な半構造化データからのデータ構造推定

研究課題名（英文）Estimating data structure embedded in semi-structured data

研究代表者
馬見塚　拓（Mamitsuka, Hiroshi）

京都大学・化学研究所・教授

研究者番号：００３４６１０７

交付決定額（研究期間全体）：（直接経費）　　13,800,000円

研究成果の概要（和文）：本研究では、グラフやネットワークに代表される半構造化データの様々な問題設定に対する解決手法を構築することを目指してきた。特に、ノードとリンクで表されるグラフにおいて、ノードに付けられたラベルに関する「ラベル伝搬」、さらに、リンクを予測する「リンク予測」という二つの問題に着目した。さらに、手法構築のみならず実データへの適用及び有効性実証を行った。この応用においては、特に生命科学におけるグラフデータ等での検証を行った。

研究成果の概要（英文）：The objective of this research is to build machine learning solutions for a variety of problems of semi-structured data, particularly graphs and networks. Particular problem focus was "label propagation" and "link prediction". We have not only built machine learning techniques but also applied our technique to real data, particularly those in life sciences.

研究分野：　生命情報科学

キーワード：　知識発見とデータマイニング

１．研究開始当初の背景

与えられたデータから内在する規則やパタンを抽出する「機械学習」あるいは「データマイニング」と呼ばれる技術は、1970 年代あるいは 1980 年代から研究が開始され、1990 年代に大きく発展し、現在さらに深耕を見せている技術分野である。この分野で考えられてきたデータは、一つ一つの事例が一定の数の変数からなる特徴量を持つ、行を事例、列を特徴とした表で与えられてきた。また、このようなデータ（構造化データ）に対する機械学習技術は既に成熟しつつあり、これらのデータに当てはまらないデータ、例えば半構造化データ、特にグラフやネットワークに対する機械学習技術の進展が待たれていた。

２．研究の目的

グラフやネットワークといった半構造化データに対する機械学習技術の構築を目的とする。通常の表のデータに対する機械学習の問題設定は、主に「分類」と「クラスタリング」の２つに分けられる。分類とは、事例にクラスラベルがあり、特徴量を用いて事例をラベル通りに分ける規則を見つけることを指す。一方、クラスタリングは、事例にクラスラベルはなく、事例の特徴量を用いてグループ化することを指す。しかし、半構造化データには、これら２つの設定のみならず、他の設定を考えることができることが、半構造化データのユニークな点である。本研究では、そのようなグラフやネットワーク特有の問題設定に対する解決手法を構築することが目的である。グラフはノード集合とノードを結合するリンクからなる。本研究で考慮するグラフ特有の問題設定の一つは、クラスラベルがすべての事例（ノード）に与えられず、かつリンクを使ってラベル未知のノードのクラスラベルを予測するラベル伝搬と呼ばれる問題設定である。もう一つの設定は、リンクの有無を予測するリンク予測と呼ばれる問題である。従って、これら２つの問題に対する解決手法を構築すると同時に、実際のデータに適用し、成果を挙げることを目的とする。

３．研究の方法

考えられる手法は主に３種類に大別できる：１）行列分解のように目的関数を設定し、正則化項とともに最適化手法を求める、２）カーネル関数による手法、３）確率モデルの最適化による手法。本研究では、特に、１）と２）に着目し、手法構築を推進するとともに可能ならば、これらを統合する手法の構築を目論む。

４．研究成果

ラベル伝搬においては、上記１）の手法に基づき、複数グラフからの新規ラベル伝搬手法を構築し、既存手法に対する優位性を検証した（下記雑誌論文＃１４）。リンク予測では、上記２の手法に基づき、与えられたデータのみからカーネル関数で内在する構造を推定することにより、従来手法、特に確率モデルによる手法をはるかに凌駕する手法を構築した（下記雑誌論文＃２１）。これらの成果を生命科学データに応用し得られた成果、並びに、研究遂行上得られた様々な成果は、30 件の雑誌論文と、4 件の学会発表論文としてまとめられている。これら研究を遂行により目的とは別に得られた成果の例として、グラフの性質を把握する上で重要なグラフラプラシアンを考察することにより、２つのグラフの類似性を測る新しい基準を提案した（下記雑誌論文＃１７）。また、さらにグラフを分割する際の基準ともできることを示した（下記雑誌論文＃９）。

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　30　件）

すべて査読あり

1. Zhou, J., Shui, Y., Peng, S., Li, X., Mamitsuka, H. and Zhu, S., MeSHSim: An R/Bioconductor Package for Measuring Semantic Similarity over MeSH Headings and MEDLINE Documents *Journal of Bioinformatics and Computational Biology*, **13** (6), 1542002 (2015). DOI: http://dx.doi.org/10.1142/S0219720015420020,
2. Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka H. and Zhu S., MeSHLabeler: Improving the Accuracy of Large-scale MeSH indexing by Integrating Diverse Evidence. *Bioinformatics* **31** (12) (*Proceedings of the 23rd International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2015)*), i339-i347 (2015). DOI: http://dx.doi.org/10.1093/bioinformatics/btv237

3. Baba, H, Takahara, J. and Mamitsuka, H., In Silico Predictions of Human Skin Permeability using Nonlinear Quantitative Structure-Property Relationship Models. *Pharmaceutical Research*, **32** (7), 2360-2371 (2015). DOI: http://dx.doi.org/10.1007/s11095-015-1629-y
4. Shiga, M. and Mamitsuka, H., Non-negative Matrix Factorization with Auxiliary Information on Overlapping Groups.*IEEE Transactions on Knowledge and Data Engineering*, **27** (6), 1615-1628 (2015). DOI: http://dx.doi.org/10.1109/TKDE.2014.2373361
5. Wang, B., Chen, X., Mamitsuka, H. and Zhu, S., BMExpert: Mining MEDLINE for Finding Experts in Biomedical Domains Based on Language Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12** (6), 1286-1294 (2015). DOI: http://dx.doi.org/10.1109/TCBB.2015.2430338
6. Yotsukura, S. and Mamitsuka, H., Evaluation of Serum-based Cancer Biomarkers: A Brief Review from a Clinical and Computational Viewpoint. *Critical Reviews in Oncology/Hematology*, **93** (2), 103-115 (2015). DOI: http://dx.doi.org/10.1016/j.critrevonc.2014.10.002
7. Mohamed, A., Hancock, T., Nguyen, C. H. and Mamitsuka, H., NetPathMiner:R/Bioconductor Package for Network Path Mining through Gene Expression. *Bioinformatics*, **30** (21), 3139-3141 (2014). DOI: http://dx.doi.org/10.1093/bioinformatics/btu501
8. Kayano, M., Shiga, M. and Mamitsuka, H., Detecting Differentially Coexpressed Genes from Labeled Expression Data:A Brief Review. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **11** (1), 154-167 (2014). DOI: http://dx.doi.org/10.1109/TCBB.2013.2297921
9. Nguyen, C. H., Wicker, N. and Mamitsuka, H., Selecting Graph Cut Solutions via Global Graph Similarity. *IEEE Transactions on Neural Networks and Learning Systems*, **25** (7), 1407-1412 (2014). DOI: http://dx.doi.org/10.1109/TNNLS.2013.2292975
10. Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S., Similarity-based Machine Learning Methods for Predicting Drug-target Interactions: A Brief Review. *Briefings in Bioinformatics* **15** (5), 737-747 (2014). DOI: http://dx.doi.org/10.1093/bib/bbt056
11. Takahashi, K., Takigawa, I. and Mamitsuka, H., SiBIC: A Web Server for Generating Gene Set Networks Based on Biclusters Obtained by Maximal Frequent Itemset Mining. *PLoS One*, 8(12), e82890 (2013). DOI: http://dx.doi.org/10.1371/journal.pone.0082890
12. Karasuyama, M. and Mamitsuka, H., Multiple Graph Label Propagation by

Sparse Integration. *IEEE Transactions on Neural Networks and Learning Systems*, **24** (12), 1999-2012 (2013). DOI: http://dx.doi.org/10.1109/TNNLS.2013.2271327

13. Shiga, M. and Mamitsuka, H., Variational Bayes Co-clustering with Auxiliary Information. *Proceedings of the 4th MultiClust Workshop on Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering (MultiClust 2013)*, Article No. 5 (2013). DOI: http://dx.doi.org/10.1145/2501006.2501012
14. Takigawa, I., Tsuda, K. and Mamitsuka, H., An *In Silico* Model for Interpreting Polypharmacology in Drug-Target Networks. *In Silico Models for Drug Discovery*, Methods in Molecular Biology, **993**, Chapter 5, 67-80 (2013). DOI: http://dx.doi.org/10.1007/978-1-62703-342-8_5
15. Nakamura, A., Saito, T., Takigawa, I., Kudo, M. and Mamitsuka, H., Fast Algorithms for Finding a Minimum Repetition Representation of Strings and Trees. *Discrete Applied Mathematics*, **161** (10-11), 1556-1575 (2013). DOI: http://dx.doi.org/10.1016/j.dam.2012.12.013
16. Gu, J., Feng, W., Zeng, J., Mamitsuka, H. and Zhu, S., Efficient Semi-supervised MEDLINE Document Clustering with MeSH Semantic and Global Content Constraints. *IEEE Transactions on Cybernetics*, **43**(4), 1265-1276 (2013). DOI: http://dx.doi.org/10.1109/TSMCB.2012.2227998
17. Wicker, N., Nguyen, C. H. and Mamitsuka, H., A New Dissimilarity Measure for Comparing Labeled Graphs. *Linear Algebra and its Applications*, **438** (5), 2331-2338 (2013). DOI: http://dx.doi.org/10.1016/j.laa.2012.10.021
18. Yamamoto, T., Nakayama, K., Hirano, H., Tomonaga, T., Ishihama, Y., Yamada, T., Kondo, T., Kodera, Y., Sato, Y., Araki, N., Mamitsuka, H. and Goshima, N., Integrated View of the Human Chromosome X-centric Proteome Project.*Journal of Proteome Research*, **12** (1), 58-61 (2013). DOI: http://dx.doi.org/10.1021/pr300844p
19. Takigawa, I. and Mamitsuka, H., Graph Mining: Procedure, Application to Drug Discovery and Recent Advance. *Drug Discovery Today*, **18** (1-2), 50-57 (2013). (Invited Review Paper) DOI: http://dx.doi.org/10.1016/j.drudis.2012.07.016
20. Hancock, T., Takigawa, I. and Mamitsuka, H., Identifying Pathways of Co-ordinated Gene Expression. *Data Mining for Systems Biology: Methods and Protocols*, Methods in Molecular Biology, **939**, Chapter 7, 69-85 (2013). DOI: http://dx.doi.org/10.1007/978-1-62703-107-3_7
21. Nguyen, C. H. and Mamitsuka, H., Latent Feature Kernels for Link Prediction on Sparse Graphs. *IEEE Transactions on Neural Networks and*

Learning Systems, **23** (11), 1793-1804 (2012). DOI: http://dx.doi.org/10.1109/TNNLS.2012.2215337

22. Hancock. T. and Mamitsuka, H., Boosted Network Classifiers for Local Feature Selection. *IEEE Transactions on Neural Networks and Learning Systems*, **23** (11), 1767-1778 (2012). DOI: http://dx.doi.org/10.1109/TNNLS.2012.2214057
23. Sorimachi, H., Mamitsuka, H. and Ono, Y., Understanding the Substrate Specificity of Conventional Calpains. *Biological Chemistry*, **393** (9), 853-871 (2012). (Invited Review Paper) DOI: http://dx.doi.org/10.1515/hsz-2012-0143
24. Mamitsuka, H., Mining from Protein-Protein Interactions. *WIREs Data Mining and Knowledge Discovery* **2** (5), 400-410 (2012). (Invited Review Paper) DOI: http://dx.doi.org/10.1002/widm.1065
25. Hancock, T., Wicker, N., Takigawa, I. and Mamitsuka, H., Identifying Neighborhoods of Coordinated Gene Expression and Metabolite Profiles. *PLoS One* **7** (2), e31345 (2012). DOI: http://dx.doi.org/10.1002/widm.1065
26. Zhang, L, Chen, Y., Wong, H.-S., Zhou, S., Mamitsuka, H. and Zhu, S., TEPITOPEpan: Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. *PLoS One* **7** (2), e30483 (2012). DOI: http://dx.doi.org/10.1371/journal.pone.0030483
27. Zhang, L, Udaka, K, Mamitsuka, H. and Zhu, S., Toward More Accurate Pan-Specific MHC-Peptide Binding Prediction: A Review of Current Methods and Tools. *Briefings in Bioinformatics* **13** (3), 350-364 (2012). DOI: http://dx.doi.org/10.1093/bib/bbr060
28. duVerle, D. and Mamitsuka, H., A Review of Statistical Methods for Prediction of Proteolytic Cleavage. *Briefings in Bioinformatics* **13** (3), 337-349 (2012). DOI: http://dx.doi.org/10.1093/bib/bbr059
29. Shiga, M. and Mamitsuka, H., Efficient Semi-Supervised Learning on Locally Informative Multiple Graphs. *Pattern Recognition* **45** (3), 1035-1049 (2012). DOI: http://dx.doi.org/10.1016/j.patcog.2011.08.020
30. Shiga, M. and Mamitsuka, H., A Variational Bayesian Framework for Clustering with Multiple Graphs. *IEEE Transactions on Knowledge and Data Engineering* **24** (4), 577-590 (2012). DOI: http://dx.doi.org/10.1109/TKDE.2010.272

〔学会発表〕（計　4　件）

すべて査読あり

1. Zheng, X., Zhu, S., Gao, J. and Mamitsuka, H., Instance-wise Weighted Nonnegative Matrix Factorization for Aggregating Partitions with Locally Reliable Clusters. *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*,

4091-4097 (2015). Buenos Aires, Argentina, Jul. 2015. **採択率 28.8% (投稿 1996 件、採択 575 件)、** http://ijcai.org/papers15/Papers/IJCAI15-574.pdf

2. Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka H. and Zhu S., MeSHLabeler: Improving the Accuracy of Large-scale MeSH indexing by Integrating Diverse Evidence. *Bioinformatics* **31** (12) (*Proceedings of the 23rd International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2015)*), i339-i347 (2015). DOI: http://dx.doi.org/10.1093/bioinformatics/btv237 Dublin, Ireland, July, 2015, **採択率 19.9% (投稿 241 件、採択 48 件)**
3. Karasuyama, M. and Mamitsuka, H., Manifold-based Similarity Adaptation for Label Propagation. *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS 2013)*, 1547-1555 (2013). Lake Tahoe, NV, USA, Dec. 2013. **採択率 25.4% (投稿 1420 件、採択 350 件)** http://papers.nips.cc/paper/5001-manifold-based-similarity-adaptation-for-label-propagation
4. Zheng, X., Ding, H., Mamitsuka, H. and Zhu, S., Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions. *Proceedings of the Nineteenth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD 2013)*, 1025-1033 (2013). Chicago, IL, USA, Aug. 2013 DOI: http://dx.doi.org/10.1145/2487575.2487670 **採択率 17.4% (投稿 726 件、採択 126 件)**

〔図書〕(計　1　件)

1. Mamitsuka, H., DeLisi, C., and Kanehisa, M., Data Mining for Systems Biology: Methods and Protocols. *Methods in Molecular Biology*, **939** (2013). (Edited book) [DOI]

〔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
馬見塚　拓(Hiroshi Mamitsuka)京都大学・化学研究所・教授、研究者番号：00346107

(2)研究分担者
なし

(3)連携研究者
瀧川　一学(Ichigaku Takigawa)北海道大学・情報科学研究科・准教授、研究者番号：10374597

ティモシー　ハンコック(Timothy Hancock)京都大学・化学研究所・助教、研究者番号：80600709

志賀　元紀(Motoki Shiga)岐阜大学・工学部・助教、研究者番号：20437263

津田　宏治(Koji Tsuda)東京大学・新領域創成科学研究科・教授、研究者番号：90357517

茅野　光範(Mitsunori Kayano)帯広畜産大学・グローバルアグロメディシン研究センター・講師、研究者番号：20590095

グェン　カン　ハオ(Canh Hao Nguyen)京都大学・化学研究所・助教、研究者番号：90626889