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

令和　５ 年　 ６ 月 ２３ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１４３４

研究課題名（和文）Machine Learning on Large Graphs

研究課題名（英文）Machine Learning on Large Graphs

研究代表者
ＮＧＵＹＥＮ　Ｃａｎｈ・Ｈａｏ（Nguyen, Canh Hao）

京都大学・化学研究所・講師

研究者番号：９０６２６８８９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：私たちはこのプロジェクトでいくつかの理論的および応用的な結果を達成しました。
理論的には、グラフの拡張であるハイパーグラフ上にすべての機械学習手法の基礎を築きました。 また、グラフ情報を活用することで、グラフに関する学習を、分子、スペクトル、および他の分子との相互作用を含むより複雑なアプリケーションに適用することもできます。

研究成果の学術的意義や社会的意義
Our achievements help pay ways for further research into more complicated problems in the area of graphs, hypergraphs and application on molecular learning. This may contribute to further research and development in biomedical applications.

研究成果の概要（英文）：We have achieved some theoretical and application results on this project. For theoretical, we laid a foundation for learning on hypergraphs, an extension of graphs. We also could apply learning on graphs to complicated applications involving molecules and its interactions with others by leveraging graph information.

研究分野：machine learning

キーワード：machine learning　Graph analysis　bioinformatics

１．研究開始当初の背景

Graph is a common way to represent relations or structures in many data analysis problems. It has applications in both theory and application of machine learning. It is an essential tool for modeling data from high-dimensional complicated distributions with manifold learning. It is important for learning with labeled and non-labeled data, a realistic scenario in many applications. It helps to model relationships in network analysis in social networks and scientific networks. In biology, chemistry and medicines, it is used to model molecules as well as their interactions.

Previous work on graph analysis are for successful for many machine learning problems. However, there are still challenges in extracting the right information from graphs, especially when the graph is too large, when the graph has complicated semantics that do not fit into usual interpretations of graphs and when more generalized versions of graphs are required.

２．研究の目的

The overall objective of the project is to extract the right semantics of structures in data in form of graphs or more complicated hyper-graphs for machine learning applications. Usual methods are for a particular semantics and do not work when the graphs have different interpretations. A main part is to find ways to extract information from different types of graphs, such as similarity or non-similarity graphs, from different applications. One of such case is to extract information from a generalized form of graph, called hypergraph. We wish to provide a principled way of extracting information from hypergraphs a provide many methods for different semantics in the hypergraphs. We also wish to extract information of graphs in complicated scenarios such as there are graphs among graphs. It is the case of biomedical applications in which molecules are graphs themselves and their interactions with other biological entities in cells form another graphs.

３．研究の方法

For the generalized form of graph called hypergraph, we have derived a uniform formulation of learning on hypergraphs. We have shown that the formulation is general enough explain the reasoning behind all available methods for hypergraphs. Moreover, it also help to create many different new methods to extract different kinds of information for many semantics of the hypergraphs. It serves as a tool not to only analyzes other method, but also to create new methods for hypergraph analysis.

type	T	t	$s(f_i, f_j)$	$sh(f)$
graph	$\sum$	.	$(f_i - f_j)^2$	$f^T Lf$: graph Laplacian [31]
graph	$\sum$	.	$\lvert f_i - f_j \rvert^p$	$< f, \Delta_p f >$: p-Laplacian [16]
graph	$\sum$	.	$\lvert f_i - f_j \rvert^{p \to \infty}$	Lipschitz extension (lex-minimizer) [22]
graph	max	.	$\lvert f_i - f_j \rvert$	Lipschitz extension (inf-minimizer) [22]
hypergraph	$\sum$	$\sum$	$(f_i - f_j)^2$	$f^T Lf$ of clique/star expansion [24]
hypergraph	$\sum$	$\sum$	$\lvert f_i - f_j \rvert$	clique expansion + 1-Laplacian
hypergraph	$\sum$	max	$\lvert f_i - f_j \rvert$	total variation [25]
hypergraph	max	max	$\lvert f_i - f_j \rvert$	inf-minimizer + star/clique expansion
hypergraph	max	$\sum$	$\lvert f_i - f_j \rvert$	max hyperedge smoothness
hypergraph	any	$(\sum \cdot)^{\frac{1}{p}}$	$\lvert f_i - f_j \rvert^p$	within-hyperedge l_p norm
hypergraph	$(\sum \cdot)^{\frac{1}{p}}$	any	$\lvert f_i - f_j \rvert^p$	between-hyperedge l_p norm

Figure 1. Our formulation leads to a unified view of methods on hypergraphs. It helps to analyze previously proposed methods (the ones with references in the table) as well as to create new ones (the ones without references in the table) with desirable properties.

For more difficult setting of graphs, especially graphs on graphs, we use a general method of graph neural networks to learn the semantics of the graphs. This is the case in many biochemical application. It helps to model molecules of interest, such as metabolites or drugs. While molecules are typically models by graph neural networks, it is more difficult when modelling their interaction with other biological components. One such case is to model molecules in interaction with spectral signals from mass spectrometry. We have derived an effective method to jointly model molecule graphs of metabolites as well as spectral data in a single framework.

4．研究成果

For the first part, we have found that our formulation led to useful hypergraph analysis methods. We could create sparse learning methods on hypergraphs, leading to sparse models, which are the models with small number of parameters that are optimally obtained. These methods show good prediction performance on classification problem on hypergraphs.

In the second part of application in metabolite identification, our method was able to extract information from molecular graphs that are useful to link to spectral information. This leads to a fast and highly accurate method that could predict metabolites of a spectrum from experiments using mass spectrometry.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　6件／うちオープンアクセス　0件）

１．著者名 Nguyen Dai Hai、Nguyen Canh Hao、Mamitsuka Hiroshi	４．巻 110
２．論文標題 Learning subtree pattern importance for Weisfeiler-Lehman based graph kernels	５．発行年 2021年
３．雑誌名 Machine Learning	６．最初と最後の頁 1587-1607
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-021-05991-y	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Nguyen Canh Hao	４．巻 29
２．論文標題 Structured Learning in Biological Domain	５．発行年 2020年
３．雑誌名 Journal of Systems Science and Systems Engineering	６．最初と最後の頁 440～453
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11518-020-5461-5	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Nguyen Hao Canh、Mamitsuka Hiroshi	４．巻 1
２．論文標題 Learning on Hypergraphs with Sparsity	５．発行年 2020年
３．雑誌名 IEEE Transactions on Pattern Analysis and Machine Intelligence	６．最初と最後の頁 1-1
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TPAMI.2020.2974746	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Nguyen Duc Anh、Nguyen Canh Hao、Mamitsuka Hiroshi	４．巻 1
２．論文標題 A survey on adverse drug reaction studies: data, tasks and machine learning methods	５．発行年 2019年
３．雑誌名 Briefings in Bioinformatics	６．最初と最後の頁 1-1
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bib/bbz140	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Nguyen Dai Hai、Nguyen Canh Hao、Mamitsuka Hiroshi	4．巻 34
2．論文標題 SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra	5．発行年 2018年
3．雑誌名 Bioinformatics	6．最初と最後の頁 i323～i332
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/bty252	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Nguyen Dai Hai、Nguyen Canh Hao、Mamitsuka Hiroshi	4．巻 35
2．論文標題 ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra	5．発行年 2019年
3．雑誌名 Bioinformatics	6．最初と最後の頁 i164~i172
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btz319	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

1．発表者名 Sun Lu、Nguyen Canh Hao、Mamitsuka Hiroshi
2．発表標題 Multiplicative Sparse Feature Decomposition for Efficient Multi-View Multi-Task Learning
3．学会等名 Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)（国際学会）
4．発表年 2019年

1．発表者名 Sun Lu、Nguyen Canh Hao、Mamitsuka Hiroshi
2．発表標題 Fast and Robust Multi-View Multi-Task Learning via Group Sparsity
3．学会等名 Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)（国際学会）
4．発表年 2019年

〔図書〕　計1件

１．著者名 Hirosato Seki, Canh Hao Nguyen, Van-Nam Huynh, Masahiro Inuiguchi	４．発行年 2019年
２．出版社 Springer	５．総ページ数 444
３．書名 Integrated Uncertainty in Knowledge Modelling and Decision Making 7th International Symposium, IUKM 2019, Nara, Japan	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関