科学研究費助成事業 研究成果報告書

平成 29 年 6 月 1 2 日現在

機関番号: 37113

研究種目: 基盤研究(B)(一般)

研究期間: 2014~2016

課題番号: 26280087

研究課題名(和文)形式グラフ体系の機械学習とグラフマイニングに関する研究

研究課題名(英文)Learning of formal graph systems and its application to graph mining

研究代表者

正代 隆義 (Shoudai, Takayoshi)

九州国際大学・現代ビジネス学部・教授

研究者番号:50226304

交付決定額(研究期間全体):(直接経費) 5,100,000円

研究成果の概要(和文):形式グラフ体系(Formal Graph System, FGS)は、Uchidaら(1994)によって導入されたグラフを対象とする一種の論理プログラムである。本研究課題では、3つの計算論的学習モデル、すなわち正例からの多項式時間帰納推論モデル、質問学習モデル、PAC学習モデルを用いて、グラフパターンと形式グラフ体系の多項式時間学習可能性を議論した。特に、形式グラフ体系がPAC学習可能であるための十分条件を、論理プログラムの各種パラメータを用いて明らかにした。また、部分グラフの分布情報を利用する分布学習アプローチにより形式グラフ体系の一部のクラスが多項式時間質問学習可能であることを示した。

研究成果の概要(英文):Formal graph system (FGS) is a logic program that deals with term graphs instead of the terms of first-order predicate logic. In this research, we have studied the polynomial-time learnability of graph patterns and formal graph systems in the frameworks of the PAC-learning, query-learning, and inductive inference models. In particular, we introduced the hierarchy of FGS that is defined by some numerical graph invariants and the parameters of logic programs. We showed that there exist formal graph systems in the hierarchy that are learnable in the PAC framework. Furthermore, we described how distributional learning techniques are applied to formal graph system (FGS) languages. We showed that the regular FGS languages of bounded degree with the 1-finite context property (1-FCP) and bounded treewidth property are learned from positive data and membership queries.

研究分野:計算機科学

キーワード: グラフ文法 グラフパターン 形式体系 計算論的学習 機械学習 帰納推論 グラフマイニング グラフアルゴリズム

1.研究開始当初の背景

(1) グラフマイニングの研究背景

このようなグラフデータを対象とするデータマイニング、すなわちグラフマイニングにおける課題のひとして、頻出部分グラフマイニングがある。これはグラフデータに頻出するローカルな共通性を抽出するものであり、Yanと Han(2002)をはじめとしても良く研究されている。グラフを入力とをる計算問題では、グラフ特有の組合せ爆発を避けるためのアプローチとして、グラフ理論的グラフテータの頻出連結部分グラフマイニングアルゴリズムを提案した。

(2) グラフパターンとその表現力の研究

グラフデータから魅力的な構造的知識を発見できるか否かは、グラフパターンのデザイン次第である。頻出部分グラフマイニングで得られる知識には、複数の頻出部分グラフがどのように配置されているか等のグローバルな構造が表現されない。

そこで、我々は、グラフデータに対する魅 力的な知識表現とその知識発見技術が重要 との認識にたち、理論と実用の両面から表現 力豊かなグラフパターンのデザインに取り 組んで来た。これまでに我々は、基本的なグ ラフパターンとして、超辺置換(Hyperedge Replacement) によるグラフパターン「項グ ラフパターン」と辺縮約(Edge Contraction) によるグラフパターン「グラフ縮約パター ン」を提案し、そのグラフ言語の多項式時間 機械学習可能性を解明した。頻出部分グラフ がローカルな共通性を表現するのに対して、 我々が定義した項グラフパターン及びグラ フ縮約パターンは複数の頻出部分グラフと その繋がり具合を同時に表現することが出 来るグローバルなグラフパターンである。

これまでに着目されてきた外平面的グラフのようなグラフ理論的グラフクラスに基づくグラフパターンクラスの設計は、グラフ理論で培われた豊かな数学的理論の結果が使えるという意味で魅力的である。しかし一

方で、グラフ理論的な枠組みにとらわれるので、細かなグラフパターンの設計が難しい。

2.研究の目的

形式グラフ体系(Formal Graph System, FGS)は、Uchida ら(1994)によって導入されたグラフを対象とする一種の論理プログラムである。形式グラフ体系によって生成される言語を形式グラフ体系言語と呼ぶ。本研究課題では、形式グラフ体系言語の機械学習可能性を探求する。木、順序木、外平面的グラフのクラスは形式グラフ言語であることから、今までのグラフパターン設計手法を統一的に扱え、さらに形式グラフ体系の論理プログラム的な特徴を深く掘り下げることで、対象とするグラフデータに相応しいグラフパターンクラス設計が可能である。

これまでに我々は、グラフ理論的グラフクラスに基づくグラフパターンの多項式時間機械学習可能性を解明して来た。しかし一方で、グラフ理論的な枠組みでは、実データに適した細かなグラフパターンの設計が難りいことも明らかになった。そこで本研究課題では、グラフ理論的グラフクラスに囚われない形式グラフ体系による新しいグラフパターンの機械学習理論の構築と、それをベースとする高速なグラフマイニングシステムの開発を目的とし、構造的知識獲得の基盤技術確立を目指した。

3.研究の方法

(1) グラフ理論的グラフクラスによるグラ フパターン学習理論

これまでに我々は、グラフパターン照合を 多項式時間で行うことができるグラフパターンの部分クラスとして、グラフ理論的グラフクラスに基づくグラフパターンを提取できる可式時間で行うフパターンのラスを定義したグラフパターンクラスを定義したグラフパターンクラスとして、区間グラフ、外平面的グラフ、Cograph、TTSPグラフ、部分 k-木がある。これらは、いずれもグラフ理論におけるグラフの不変量を使って、実用的な時間で適したののである。本研究課題では、これらの結果の知識と経験を活かし、グラフ理論的グラフがゴリズムの設計と解析を行った。

(2) 形式グラフ言語の学習アルゴリズムの 設計と計算論的学習理論の評価

形式グラフ体系によるグラフクラスが、グラフ理論的グラフクラスの階層構造に置いて、どのように位置づけられるか、特に、グラフ同型判定性問題等、グラフマイニングにおける重要な計算問題に焦点をあてた計算量理論的階層構造の解明を行うことが必要である。Okadaら(2007)は、形式グラフ体系により生成されたグラフパターンの機械学習理論を扱った。これは、計算論的学習理論

の主要な学習モデルのひとつである質問学習について、形式グラフ体系により生成されるグラフ言語の学習可能性を議論した結果である。この結果でのアイデアを基に、形式グラフ体系によるグラフパターンクラスを統一的に扱える効率の良いグラフパターン学習アルゴリズムの設計を行った。

正例からの多項式時間帰納推論は計算論的学習理論において研究の中心となる学習モデルのひとつである。正例からの多項式時間帰納推論では、学習ターゲットは仮ターゲットを特殊化する方向で多項式時間学習アルゴリズムを設計することが多い。一方で、質問学習では一般化する方向で設計することが多い。両学習モデルの特徴に十分配慮しつの、形式グラフ言語の多項式時間学習アルゴリズムの設計と解析を行った。

(3) グラフマイニングの確率化・近似化

論理プログラムの機械学習に関する限界は、我々のグラフパターンクラスにも同じかそれ以上に重要な問題となる。これには、グラフパターンクラス空間の効率の良い探索手法の確立が不可欠である正例からの帰納推論では、適切な精密化演算子を定めることが、探索手法の効率の良さに深く関係している。計算量理論的な効率の良さにこだわらず、積極的に確率的・近似的・発見的手法を取り入れ、実社会で使える技術の構築を目指した。

4. 研究成果

(1) グラフ理論的グラフクラスによるグラフパターン学習理論

順序項木パターンは、構造的変数を持つこ とが出来るラベル付き順序木構造パターン で、XML 文書や RNA の二次構造のような順 序木構造パターンを表現することが出来る。 一方、文字列におけるデータベースとしてよ く知られているタンパク質データベース PROSITE では、文字列で表されたタンパク 質の類似度を表現する手段として、文字列長 に制限を加えることが出来るパターン (PROSITE パターン)が定義されている。順 序木構造パターンとして、PROSITE パター ンのような木の高さや幅を制約するパター ンは知られていなかったが、[雑誌論文 4]で は、順序木構造パターンとして高さ制約付き 順序項木パターンのクラスを定義し、このパ ターンクラスが、構造的変数における変数ラ ベルが全て異なり、かつ構造的変数をチェー ンとして持たないとき、正例から多項式時間 帰納推論可能であることを証明した。

GとHを無向グラフとする。Gを辺縮約によってHに変換できるかを考える。ただし、次のような条件を設定する。Hの頂点集合の部分集合Uを次の条件を満たすために指定する。HのUに属さない頂点Vは、Vへ縮約されるGの頂点数がちょうどI 個となる頂点である。[雑誌論文 15]では、グラフHの頂

点のうち、独立かつ次数 2 の頂点を U の頂点と指定すると、H が部分 k-木 (k:固定)である場合に、G から H への変換が可能かを問う問題が多項式時間計算可能であることを示した。これは、部分 k-木に基づくグラフ縮約パターンクラスが正例から多項式時間帰納推論可能であるための十分条件の一部である。

[雑誌論文 11]では順序項木パターンの多項式時間アルゴリズムを提案した。本アルゴリズムは、頂点数 N の順序木と頂点数 n の順序木と頂点数 n の順序木と頂点数 n の順序本とであるか否かを O(nN)時間で判定する。さらするか否かを O(nN)時間で判定する。がらに[雑誌論文 1]では、部分 k-木に基づく項グラフパターンクラスに対する多項式時間照合アルゴリズムを提案した。部分 k-木に列から多項式時間帰納推論可能であることを示すことは今後の重要な課題である。

文字列圧縮手法の1つであるLZ法を順序木の構造圧縮に応用し木構造データを圧縮して得られる内部辞書を持った順序木を圧縮項木と呼ぶ。これは、順序項木パターンに変数への内部リンクを保持させたものとみなせる。[雑誌論文 5]では、この圧縮項木の順序木構造を、超簡潔表現の1つであるDFUDS (Depth-First Unary Degree Sequence)で表現し、陽に圧縮項木を解凍することなく頻出するパスを全て枚挙する高速なアルゴリズムを提案した。[雑誌論文 3]では、順序木集合が与えられたとき、頻出で特徴的な順序項木パターンを漏れ無く列挙するアルゴリズムを提案した。

(2) 形式グラフ言語の学習アルゴリズムの 設計と計算論的学習理論の評価

[雑誌論文 13]では、文脈決定正則形式グラフ体系を形式グラフ体系のサブクラスとして導入し、そのクラスにより生成されるグラフ言語の全体が、多項式時間で教師付き学習(MAT)可能であることを証明した。

また、形式グラフ体系により設計されたグラフパターンクラスの PAC 学習可能性についても論じた。PAC 学習とは、学習アルゴリズムの出す仮説に誤差を許す学習モデルで、現実世界のデータを扱う際、データの揺らぎを許容できるので都合がよい。[雑誌論文 14]では、形式グラフ体系が PAC 学習可能であるための十分条件を証明した。

語句の分布情報を利用する分布学習を形式グラフ言語に一般化した。[雑誌論文 2]では、形式グラフ体系が、(1) 1-有限文脈性を有すること、(2) 正規であること、(3) 次数と木幅が有界であるグラフだけを生成すること、以上の3条件を満たすならば、所属性質問を用いて多項式更新時間で正例から帰納推論可能であることを証明した。ここで、1-有限文脈性とは、形式グラフ体系の各述語に対して、その述語が定義するグラフ言語が一変数グラフパターン言語であることをい

う。また、形式グラフ体系が正規であるとは、 その形式グラフ体系が超辺置換文法と等価 であることをいう。本定理は、分布学習アプ ローチによる形式グラフ体系言語のグラフ マイニングの理論的基盤となる結果である。

(3) グラフマイニングの確率化・近似化 TTSP グラフはネットワークのデータモデルとして用いられている。[学会発表 2]では、グラフパターン集合を個体とする進化的計算を用いて、正事例と負事例の TTSP グラフから特徴的な複合的グラフ構造パターンを獲得する手法を提案した。

多くの化合物は外平面的グラフの構造を 持っている。[雑誌論文 6,8-10]は、進化的計 算が特徴的な外平面的グラフパターン発見 に効果的であることを示す結果である。

[学会発表 3]では、グラフマイニングアルゴリズムの実装及び視覚化を行った。また、外平面的グラフパターン発見アルゴリズムの実装と GUI 開発を行い、高校生向け公開講座等で実演を行った。

実データを扱ったデータマイニング研究として[雑誌論文 7]では、トラヒックデータ中のパケット間距離とパケットクラス分類に基づいて作成した重み付き有向グラフに対して、グラフカット半教師あり学習を適用する手法を提案した。本論文では、計算量理論的な効率の良さにこだわらず、確率的・近似的・発見的手法を取り入れ、実社会で支える技術の構築を目指した。

5 . 主な発表論文等

[雑誌論文](計15件)

T. Shoudai and T. Yamada, A Polynomial Time Pattern Matching Algorithm on Graph Patterns of Bounded Treewidth, IEICE Trans. Fundamentals, 查読有, Vol.E100-A, No.9, 2017, to appear.

T. Shoudai, S. Matsumoto, and Y. Suzuki, Distributional Learning of Regular Formal Graph System of Bounded Degree, Proc. 26th Int. Conf. on Inductive Logic Programming (ILP2016), Springer, Lecture Notes in Artificial Intelligence, 查読有, 2017, to appear.

T. Miyahara, <u>Y. Suzuki</u>, <u>T. Shoudai</u>, <u>T. Uchida</u>, and T. Kuboyama, Enumeration of Maximally Frequent Ordered Tree Patterns with Wildcards for Edge Labels, IPSJ Trans. Mathematical Modeling and its Applications, 查読有, 2017, to appear.

T. Shoudai, K. Aikoh, Y. Suzuki, S. Matsumoto, T. Miyahara, and T. Uchida, Polynomial Time Inductive Inference of Languages of Ordered Term Tree Patterns with Height-Constrained Variables from Positive Data, IEICE Trans. Fundamentals, 查読有, Vol.E100, No.3, 2017, 785-802.

DOI: 10.1587/transfun.E100.A.785

T. Horibe, Y. Itokawa, <u>T. Uchida, Y. Suzuki</u>, and T. Miyahara, Algorithm for Enumerating all Frequent Paths from Structurally Compressed Tree-Structured Data, Proc. Int. MultiConf. of Engineers and Computer Scientists 2017, 查読有, Vol I, 2017, 69-74.

http://www.iaeng.org/publication/IME CS2017/IMECS2017 pp69-74.pdf

F. Tokuhara, T. Miyahara, Y. Suzuki, T. Uchida, and T. Kuboyama, Using Canonical Representations of Block Tree Patterns in Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns, Proc. 9th Int. Workshop on Computational Intelligence and Applications (IWCIA2016), 查読有, 2016, 93-99. DOI: 10.1109/IWCIA.2016.7805755

T. Shoudai, H. Murai, and A. Okamoto, A Semi-Supervised Data Screening for Network Traffic Data Using Graph Min-Cuts, IPSJ Trans. Mathematical Modeling and its Applications, 查読有, Vol.9, No. 2, 2016, 49-60.

https://ipsj.ixsq.nii.ac.jp/ej/?action=pa ges view main&active action=reposi tory view main item detail&item id =174206&item no=1&page id=13&bl ock id=8

F. Tokuhara, T. Miyahara, <u>Y. Suzuki</u>, <u>T. Uchida</u>, and T. Kuboyama, Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns by Genetic Programming using Label Information, Proc. 5th IIAI Int. Congress on Advanced Applied Informatics (IIAI-AAI2016), 查読有, 203-210, 2016.

Y. Ouchiyama, T. Miyahara, <u>Y. Suzuki</u>, <u>T. Uchida</u>, T. Kuboyama, and F. Tokuhara, Acquisition of

DOI: 10.1109/IIAI-AAI.2016.212

Characteristic Block Preserving Outerplanar Graph Patterns from Positive and Negative Data using Genetic Programming and Tree Representation of Graph Patterns, Proc. 8th Int. Workshop on Computational Intelligence and Applications (IWCIA 2015), 查読有, 2015, 95-101.

DOI: 10.1109/IWCIA.2015.7449469

S. Tani, T. Miyahara, <u>Y. Suzuki</u>, and <u>T. Uchida</u>, Acquisition of Multiple Tree Structured Patterns by an Evolutionary Method using Sets of Tag Tree Patterns as Individuals, Proc. 4th IIAI Int. Congress on Advanced Applied Informatics (IIAI-AAI2015), 查 読 有 , 2015, 213-218.

DOI: 10.1109/IIAI-AAI.2015.271

Y. Suzuki, T. Shoudai, T. Uchida, and T. Miyahara, An Efficient Pattern Matching Algorithm for Ordered Term Tree Patterns, IEICE Trans. Fundamentals, 查読有, Vol.E98-A, No.6, 2015, 1197-1211.

DOI: 10.1587/transfun.E98.A.1197

S. Nakai, T. Miyahara, <u>Y. Suzuki</u>, T. Kuboyama, and <u>T. Uchida</u>, Acquisition of Characteristic Sets of Tree Patterns with VLDC's Using Genetic Programming and Edit Distance, Proc. 7th Int. Workshop on Computational Intelligence and Applications (IWCIA2014), 查読有, 2014, 113-118.

DOI: 10.1109/IWCIA.2014.6988088

S. Hara and <u>T. Shoudai</u>. Polynomial Time MAT Learning of C-Deterministic Regular Formal Graph Systems. Proc. 3rd Int. Congress on Advanced Applied Informatics (IIAI-AAI2014), 查読有, 2014, 204-211.

DOI: 10.1109/IIAI-AAI.2015.51

正代 隆義, 多項式時間 PAC 学習可能 な木幅定数形式グラフ体系について. 九州国際大学教養研究, 査読無, 第 21 巻, 第 1-2 合併号, 2014, 1-26.

https://kiu.repo.nii.ac.jp/?action=page s view main&active action=reposito ry view main item detail&item id= 514&item no=1&page id=25&block i d=33 T. Yamada and <u>T. Shoudai</u>, Graph Contraction Pattern Matching for Graphs of Bounded Treewidth, Latest Advances in Inductive Logic Programming, Imperial College Press, 查読無, 173-180, 2014.

DOI: 10.1142/9781783265091 0018

[学会発表](計20件)

松本 哲志, 正代 隆義, 鈴木 祐介, グラフ構造の分布情報を用いた有界な次数と木幅の文脈自由グラフ言語の多項式時間機械学習, 情報処理学会第79回全国大会, 5A-06, 2017年3月17日, 名古屋大学東山キャンパス.

山縣 佑貴, 宮原 哲浩, <u>鈴木 祐介</u>, <u>内</u> 田 智之, 久保山 哲二, TTSP グラフパターン集合を個体とする進化的手法による複合的グラフ構造パターンの獲得,情報処理学会九州支部火の国情報シンポジウム 2017, 2017 年 3 月 2 日, 鹿児島大学工学部.

延 拓哉, <u>鈴木 祐介</u>, 内田 智之, 宮原哲浩, 化合物データからの特徴的な外平面的グラフパターン発見アルゴリズムの実装と GUI 開発, 情報処理学会九州支部火の国情報シンポジウム 2017, 2017年3月1日, 鹿児島大学工学部.

T. Miyahara, Y. Suzuki, T. Shoudai, T. Uchida, and T. Kuboyama, Enumeration of Maximally Frequent Ordered Tree Patterns with Wildcards for Edge Labels, 第112回数理モデル化と問題解決研究発表会, 2017年2月27日, 盛岡市繋温泉.

李 起春, <u>鈴木 祐介</u>, <u>内田 智之</u>, 宮原 哲浩, 2-限定的木置換文法に対する非終端記号に関する所属性質問を用いた質問学習, 平成 28 年度電気・情報関連学会中国支部連合大会, 2016 年 10 月 22 日, 広島大学東広島キャンパス.

木下 智貴, <u>鈴木 祐介</u>, <u>内田 智之</u>, 宮原 哲浩, 隣接した変数を持つ順序グラフパターンに対するマッチングアルゴリズム, 平成 28 年度電気・情報関連学会中国支部連合大会, 2016 年 10 月 22 日, 広島大学東広島キャンパス.

大内山 祐斗, 宮原 哲浩, <u>鈴木 祐介</u>, 内田 智之, 久保山 哲二, 徳原 史也, グラフパターンの木表現を用いた遺伝 的プログラミングによる特徴的なブロ

ック保存型外平面的グラフパターンの 獲得,2016 人工知能学会全国大会, 2016年6月8日,北九州国際会議場.

徳原 史也, 宮原 哲浩, <u>鈴木 祐介</u>, <u>内</u> 田 智之, 久保山 哲二, ラベルの情報 を利用した遺伝的プログラミングによる特徴的なブロック保存型外平面的グラフパターンの獲得, 火の国情報シンポジウム 2016, 2016 年 3 月 2 日, 宮崎大学木花キャンパス

重松 卓也, 宮原 哲浩, <u>鈴木 祐介</u>, <u>内</u> 田 智之, タグ木パターン集合に作用する遺伝的プログラミングによる特徴的な複合的木構造パターンの獲得, 火の国情報シンポジウム 2016, 2016 年 3 月 2 日, 宮崎大学木花キャンパス.

堀部 智也, 糸川 裕子, 内田 智之, 宮原 哲浩, <u>鈴木 祐介</u>, LZ 法に基づいて 構造圧縮された順序木構造データに対 する頻出パス枚挙アルゴリズム, 火の 国情報シンポジウム 2016, 2016 年 3 月 2 日, 宮崎大学木花キャンパス.

中居 翔平, 宮原 哲浩, <u>鈴木 祐介</u>, 久保山 哲二, <u>内田 智之</u>, VLDC 木パターン集合を個体とする進化的手法による複合的木構造パターンの獲得, 2015 人工知能学会全国大会, 2015 年 5 月 31 日, 公立はこだて未来大学.

糸川 裕子, 内田 智之, 構造データからの頻出多ポート項木パターン枚挙アルゴリズム, 2015 人工知能学会全国大会, 2015年5月30日, 公立はこだて未来大学.

正代 隆義, 村井 光, 岡本 敦, A Semi-Supervised Data Screening for Network Traffic Data using Graph Min-Cuts,情報処理学会第 107 回数理モデル化と問題解決(MPS) 研究会, 2016年3月8日, 山口市湯田温泉.

正代 隆義, 文脈決定正則形式グラフ体系の多項式時間 MAT 学習, 2015年7月4日, 第13回論理と計算セミナー,福岡工業大学.

村井 光, 正代 隆義, 効果的なネット ワークインシデント検知のための半教師ありデータスクリーニング, 火の国情報シンポジウム 2015, 2015 年 3 月 5 日, 佐賀大学本庄キャンパス.

村井 光, 正代 隆義, グラフベースの 半教師あり学習によるデータスクリー ニングソフトウェア, Software in Mathematics Demonstration Track in Hakata Workshop 2015, 2015 年 2 月 6 日, 福岡市博多区.

正代 隆義, 内田 智之, 多項式時間学習可能な木幅定数グラフ言語の形式体系について, 冬の LA シンポジウム, 2015年1月30日, 京都大学数理解析研究所.

正代 隆義, 多項式時間 PAC 学習可能な木幅定数形式グラフ体系について, 第 19 回情報・統計科学シンポジウム, 2014年12月5日, 九州大学箱崎キャンパス.

T. Shoudai, S. Hara, and T. Uchida, Polynomial time learning of formal graph systems via queries, ELC Workshop on Learning Theory and Complexity collocated with 12th Int. Conf. on Grammatical Inference (ICGI 2014), 2014年9月16日, 京都大学.

日野 隆博, <u>鈴木 祐介</u>, 内田 智之, 宮原 哲浩, 順序グラフパターン言語の多項式時間帰納推論, 2014 人工知能学会全国大会, 2014 年 5 月 14 日, 愛媛県県民文化会館.

6.研究組織

(1) 研究代表者

正代 隆義 (SHOUDAI, Takayoshi) 九州国際大学・現代ビジネス学部・教授 研究者番号: 50226304

(2) 研究分担者

内田 智之 (UCHIDA, Tomoyuki) 広島市立大学・大学院情報科学研究科・准 教授

研究者番号: 70264934

松本 哲志 (MATSUMOTO, Satoshi) 東海大学・理学部・准教授 研究者番号: 30307235 (平成 28 年度より研究分担者)

鈴木 祐介 (SUZUKI, Yusuke) 広島市立大学・大学院情報科学研究科・助 教

研究者番号: 10398464 (平成 28 年度より研究分担者)