科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 １ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５００１８２

研究課題名（和文）クラス指向グラフパターン設計手法の開発とグラフマイニングへの応用

研究課題名（英文）Design and Analysis of Efficient Class-oriented Graph Mining Systems

研究代表者
正代　隆義（Shoudai, Takayoshi）

九州大学・システム情報科学研究科（研究院・准教授

研究者番号：５０２２６３０４

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：本研究課題では、グラフパターンクラスの設計手法とグラフパターンの多項式時間学習アルゴリズムを開発し、主として次の2つの結果を得た。
(1) 現実に最もよく現れるグラフ構造は木構造である。我々は、辺縮約に基づくグラフ構造の新しいパターン表現「木縮約パターン」を提案し、そのクラスが正データから多項式時間機械学習可能となる条件を示した。さらに、計算困難性・近似困難性についても議論し、木縮約パターンの限界を明らかにした。
(2) グラフ文法の一つである形式グラフ体系(FGS)に対して、高い表現力と高速な機械学習を両立させるグラフパターンクラスが形式グラフ体系のクラスとの関連で議論できることを示した。

研究成果の概要（英文）：Our object of this research is to develop an effective graph pattern designing system for efficient data mining from graph-structured data. During this research period, the following results mainly were obtained.
(1) A tree contraction pattern (TC-pattern) is an unordered tree-structured pattern common to given unordered trees, which is obtained by merging every uncommon connected substructure into one vertex by edge contraction. We show that an important subclass of TC-patterns is polynomial-time inductively inferable from positive data. Moreover, we discuss the optimization versions of the learning problems for TC-patterns, and give the conditions under which the optimization problems are hard to compute.
(2) We introduce context-deterministic regular formal graph systems (FGS) as one of the effective graph pattern designing systems, and propose a polynomial time algorithm for learning the class of context-deterministic regular FGSs in the framework of MAT learning.

研究分野：総合領域

科研費の分科・細目：情報学・知能情報学

キーワード：グラフパターン　グラフマイニング　グラフアルゴリズム　グラフ構造データ　データマイニング　機械学習　機械発見　帰納推論

１．研究開始当初の背景

(1) 薬理学の分野では化学化合物の分子構造を解明するために、原子を頂点、原子間の化学結合を辺としたグラフのデータマイニングが行われている。また、社会学におけるソーシャルネットワークの解析や、ウェブページ間のリンク構造を表すウェブグラフを利用したウェブコミュニティの抽出など、グラフとして表現したデータを直接扱うこと、及びその技術の研究が急速に行われてきている。グラフマイニングにおける主要な課題のひとつである頻出部分グラフマイニングには、gSpan [Yan と Han, 2002]の他に、AGM [Inokuchi ら., 2000], Gaston [Nijssen と Kok, 2004], FSG [Kuramochi と Karypis, 2001] などがある。またこれらのシステムを基に、より大規模なグラフマイニングのための情報論的・統計的な研究が行われている。

(2) こうしたグラフマイニングの一連の研究の中で、我々は、独自の視点のもとに、表現力に富む木構造パターンを定義した。さらに、木構造パターン言語の多項式時間機械学習の理論を展開し、計算論的学習理論に関する国際会議でその研究成果を発表してきた。さらにそういった理論面の成果が机上の空論でないことの実証として、木構造パターン言語のための多項式時間機械学習アルゴリズムによる木構造データからのデータマイニングシステムを実現してきた。

(3) 化学化合物の分子構造はその多くが外平面的グラフとよばれるグラフクラスで表現可能である。我々と同様に、グラフマイニングにおける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クラスに着目した研究に Horváth らの頻出連結部分グラフマイニングの研究がある。我々は、部分グラフよりも表現力を重視して、ブロック保存型外平面的グラフパターン(BPO グラフパターンと略す)とグラフマイニングアルゴリズムを提案し、NCI Chemical Dataset (http://cactus.nci.nih.gov/)から効率良くパターンを発見することに成功した。BPO グラフパターンは変数構造を持つグラフパターンである。BPO グラフパターンはいくつかの部分グラフ(ブロック)の繋がりを超辺置換(Hyperedge Replacement)により表現したものであり、部分グラフよりも真に表現力が大きい。化学化合物のうち外平面的グラフではないグラフは部分 k-木 (k:固定)である。化学化合物の頻出グラフパターンの全貌を明らかにするため、過去の研究で我々は、部分 k-木に基づくグラフパターンの導入とマイニングアルゴリズムの提案を行った。

２．研究の目的

(1) これまでに着目されてきた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。そこで、我々は、グラフを項として扱える FGS(Formal Graph System をグラフパターンクラスの設計に用いるための機械学習理論の構築を行うことを目標とした。文字列言語の機械学習でよく研究されている言語表現に EFS (Elementary Formal System)がある。EFS は文字列を直接扱うことのできる一種の論理プログラムで、帰納推論などの学習の能力を示すのに都合のよい体系である。FGS は、グラフを項として直接操作できる論理プログラミングシステムであり、EFS の自然なグラフ理論的拡張である。FGS はグラフで表現できる様々なオブジェクト間の関係を論理的に表現するのに適している。

(2) これまでの研究の対象としてきた木、外平面的グラフのクラスでは、計算論的学習理論に基づいたグラフマイニングアルゴリズムを提案し、それらの実装を行ってきた。本研究では、これまでの研究で培われてきた計算論的学習理論に基づくグラフマイニングのアイデアと知識が自然な形で導入できる。また、既に実装済みのプログラムをベースとするグラフマイニングシステムの構築を目指す。すなわち、本研究課題の目的は、グラフパターンクラス設計手法の確立を行い、新しいグラフマイニングシステムのソフトウェア及びハードウェア両面からの高速化とその限界を明らかにすることである。

３．研究の方法

(1) 正データもしくは完全データからの帰納推論は、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。我々は、木構造パターンクラスと外平面的グラフパターンクラスの正データからの多項式時間帰納推論可能性について、詳細に議論し、一連の成果を得た。どちらのグラフパターンクラスに対しても、グラフ理論におけるグラフの特徴量を巧妙に使うことで、実用的な時間で有用なグラフパターンを発見することに成功し、さらにそのグラフパターンクラスの正データからの多項式時間帰納推論可能性を証明してきた。我々は、この知識と経験を生かして、グラフデータベースに対するグラフパターンによる知識獲得手法の開発を行った。

(2) グラフパターンクラス設計のためには、発見されたグラフパターンから得られるメタな知識のフィードバックの仕組みが必要である。このためにはグラフパターンクラスの高速な探索アルゴリズムの設計が必要であり、これらのアルゴリズムをエンジンとして効率の良い知識の洗練化システムの実働化を行う必要がある。これに関しては、マル

チコア CPU コンピュータを用い、本格的なグラフマイニングシステムの開発を行うこととした。

(3) グラフパターンクラスの探索には、非常に多くの CPU パワーを必要とするため、コストパフォーマンスと拡張性を重視し、GPU コンピュータを用いた。高速グラフパターン照合・発見アルゴリズムの研究は、計算機シミュレーションでの効率を現実的に評価した。実際、計算量の大きいグラフパターンの発見を行うために、マルコフ連鎖モンテカルロ法などの確率的手法を積極的に導入し、アルゴリズムの規模耐久性を目指し、実働可能なことを確認している。

４．研究成果

(1) *Cograph パターン言語の多項式時間機械学習に関する研究* [雑誌論文 10]

グラフ理論的なグラフクラスである Cograph のクラスは、ある種のスケジューリング問題や索引のクラスタリングなどで用いられている。Cograph のクラスは、1 頂点からなるグラフから始めて、互いに共通部分のない 2 つのグラフの和をとる操作と、グラフの捕グラフを求める操作を繰り返し適用して得られるグラフの全体である。本研究では、Cograph のクラスをベースとする新しいグラフパターンクラスを設計し、Cograph パターンのクラスが正データから多項式時間帰納推論可能であることを証明した。このことは、一部のスケジューリングに関するデータ等、Cograph で表現することができるデータから、新しい知識を発見するための理論的基盤となる。Cograph とは限らない一般のグラフに対しては､辺を補完し Cograph に変換するいくつかの方法が知られている。これが効果的にできれば、実データに対する効率の良いグラフマイニング手法の開発が期待できる。

(2) *木縮約パターンの多項式時間機械学習とその限界に関する研究* [雑誌論文 1,3,5,9]

任意のグラフに対して、辺で結ばれた 2 頂点を 1 つの頂点に融合する操作を辺縮約と呼ぶ。我々は、過去の研究で、グラフ文法の超辺置換に基づくグラフパターン、項グラフパターンと呼ぶ、について様々な側面から多項式時間機械学習理論を展開してきた。しかし、一方で、この項グラフパターンには、頂点の次数に関するパターン表現の限界も明らかになってきた。本研究では、このような限界を克服するために、辺縮約に基づくグラフ構造の新しいパターン表現を提案した。一般のグラフに対する、このグラフパターン表現を「グラフ縮約パターン」と呼ぶ。また、木に対するグラフ縮約パターンを「木縮約パターン」と呼ぶ（図 1)。このグラフ縮約パターンに関して、研究を行い、主として次の 3 つの結果を得た。

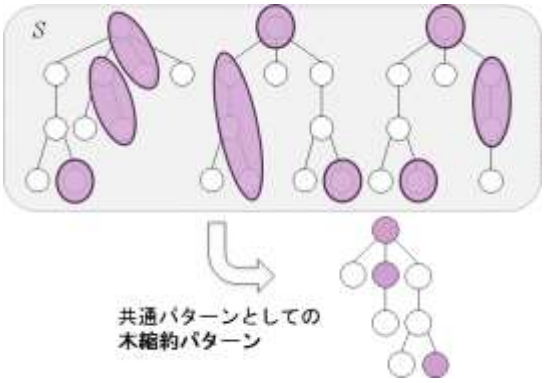

図 1木縮約パターン[雑誌論文 9]: Sに含まれる 3 つの木の色付き領域を 1 つの頂点とみなせば、下段のパターンが導かれる。

① グラフ構造パターンの照合アルゴリズムを提案し、グラフ構造パターンの木幅が定数と見なせる場合、多項式時間で動作することを示した。ほとんどの化学化合物グラフは木幅が定数とみなせることから、本アルゴリズムを用いたグラフマイニングシステムを開発することで、化学化合物における新しい知識発見が期待される。

② 扱いやすいグラフ構造データである木構造データに対するパターン表現として項木パターンがある。これは超辺置換により共通構造を柔軟に表現するが、超辺の次元次第でパターン照合問題が NP 困難になる。そこで、我々は前述のように新しい木構造パターン、木縮約パターンを導入した。木縮約パターンは頂点を変数とみなすので、項木パターンとは異なる共通構造表現が可能である。本研究は、照合アルゴリズムの時間計算量を向上させるとともに、変数頂点の次数をパラメータとする FPT アルゴリズムを提案した。また、木構造データからの効率的かつ効果的な木構造パターンマイニングを目的として、木縮約パターンの最悪多項式時間遅延列挙アルゴリズムを提案した。

③ 半構造データの重要な機械学習問題の一つとして、木構造データの二値分類問題がある。我々は、学習の高速化を目指して，木縮約パターンによる木構造データの二値分類問題に対するマルコフ連鎖モンテカルロ法を用いたアルゴリズムを提案した。そのアルゴリズムを糖鎖データに対して適用した結果、カーネル法に基づくアルゴリズムの結果には劣るが、一つのパターンで分類を行う手法としては良好な結果を得た。また、この問題に関しては、計算困難性・近似困難性についても議論し、木縮約パターンの限界を明らかにした。

(3) *時系列パターン発見のための近似ストリームアルゴリズム* [雑誌論文 6,7,11]

ストリームデータに頻出する時系列を列挙する近似ストリームアルゴリズムを提案した。提案したアルゴリズムは、時間変化するグラフデータの特徴をグラフパターンとして捉えるための基礎と成り得る。

また、グラフ系列の特徴をグラフの更新ルールとして捉えるための基盤として、ストリームデータの頻出イベント系列を列挙するストリームアルゴリズムについて研究を行った。イベント系列出現数をどのようにカウントするかは頻出イベント系列列挙の重要な鍵となる。我々は、先着優先による頻出度カウントのもとで、精度保証付き頻出イベント系列発見ストリームアルゴリズムを提案し、実データ上で提案アルゴリズムの有効性を確認した。グラフ系列に対するストリームアルゴリズムの提案は今後の重要な課題である。

(4) *形式グラフ体系のサブクラスに対する多項式時間機械学習*

形式グラフ体系(FGS)は一階述語論理の項の代わりに超グラフを扱う論理プログラムであり、グラフ文法の一つとみなすことができる。我々は、形式グラフ体系により定義されるグラフパターンで，文脈決定可能かつ多項式時間質問学習可能なクラスの一つを示した。このことは、高い表現力と高速な機械学習を両立させるグラフパターンクラスが形式グラフ体系のクラスとの関連で議論できることを示しており、今後の研究を進める上で意義深い結果である。

５．主な発表論文等

〔雑誌論文〕（計 11 件）

① Y. Okamoto and T. Shoudai, Hard Optimization Problems in Learning Tree Contraction Patterns, Applied Computing and Information Technology, Studies in Computational Intelligence, 査読有, Vol.553, 2014, 77-90
DOI: 10.1007/978-3-319-05717-0_6

② Y. Itokawa, T. Uchida, and M. Sano, An Algorithm for Enumerating All Maximal Tree Patterns Without Duplication Using Succinct Data Structure, Proc. Int. MultiConf.. Engineers and Computer Scientists 2014 (IMECS 2014), 査読有, Vol.I, 2014, 156-161
http://www.iaeng.org/IMECS2014/

③ Y. Okamoto, K. Koyanagi, T. Shoudai, and O. Maruyama, Discovery of Tree Structured Patterns Using Markov Chain Monte Carlo Method, Proc. 7th IADIS Int. Conf. Information Systems 2014, 査読有, 2014, 95-102
http://www.is-conf.org/

④ T. Hino, Y. Suzuki, T. Uchida, and T. Miyahara, Ordered Graph Patterns Which Are Polynomial Time Inductively Inferable from Positive Data, Proc. 7th IADIS Int. Conf. Information Systems 2014, 査読有, Vol.I, 2014, 263-270
http://www.is-conf.org/

⑤ Y. Okamoto and T. Shoudai, Hardness of Learning Unordered Tree Contraction Patterns, Proc. 2nd Int. Conf. Advanced Applied Informatics (IIAI-AAI2013), 査読有, 2013, 141-146
DOI: 10.1109/IIAI-AAI.2013.63

⑥ H. Tsuruta and T. Shoudai, Structure-based Data Mining and Screening for Network Traffic Data, Proc. 2nd Int. Conf. Advanced Applied Informatics (IIAI-AAI2013), 査読有, 2013, 152-157
DOI: 10.1109/IIAI-AAI.2013.78

⑦ A. Okamoto and T. Shoudai, Mining First-Come-First-Served Frequent Time Sequence Patterns in Streaming Data, Proc. IADIS Int. Conf. e-Society 2013, 査読有, 2013, 283-290
http://www.esociety-conf.org/

⑧ T. Hino, Y. Suzuki, T. Uchida, and Y. Itokawa, Polynomial Time Pattern Matching Algorithm for Ordered Graph Patterns, Lecture Notes in Artificial Intelligence, 査読有, Vol.7842, 2013, 86-101
DOI: 10.1007/978-3-642-38812-5_7

⑨ Y. Yoshimura and T. Shoudai, Learning Unordered Tree Contraction Patterns in Polynomial Time, Lecture Notes in Artificial Intelligence, 査読有, Vol.7842, 2013, 257-272
DOI: 10.1007/978-3-642-38812-5_18

⑩ Y. Yoshimura, T. Shoudai, Y. Suzuki, T. Uchida, and T. Miyahara, Polynomial Time Inductive Inference of Cograph Pattern Languages from Positive Data, Lecture Notes in Artificial Intelligence, 査読有, Vol.7207, 2012, 389-404
DOI: 10.1007/978-3-642-31951-8_32

⑪ H. Tsuruta, T. Shoudai, and J.

Takeuchi, Network Traffic Screening Using Frequent Sequential Patterns, Lecture Notes in Electrical Engineering, 査読有, Vol.110, 2012, 363-375
DOI: 10.1007/978-1-4614-1695-1_28

〔学会発表〕(計 12 件)

① 小柳 健介, 岡本 康宏, 正代 隆義, 丸山 修, マルコフ連鎖モンテカルロ法による最適木構造パターン発見手法, 2014 年電子情報通信学会総合大会, 2014 年 03 月 18 日～2014 年 03 月 21 日, 新潟県新潟市

② 佐野 元紀, 内田 智之, 糸川 裕子, 簡潔データ構造を用いた極大項木パターン枚挙アルゴリズム, 2014 年電子情報通信学会総合大会, 2014 年 03 月 18 日～2014 年 03 月 21 日, 新潟県新潟市

③ 溝口 佳寛, 田中 久治, 坂下 一生, 井口 修一, 有限オートマトンとスティッカー系に関する Coq による形式証明について, 日本数学会 2014 年度年会, 2014 年 3 月 15 日～2014 年 3 月 18 日, 学習院大学目白キャンパス

④ 小柳 健介, 岡本 康宏, 丸山 修, 正代 隆義, マルコフ連鎖モンテカルロ法の木構造パターン発見への応用, 2013 年度夏の LA シンポジウム, 2013 年 07 月 16 日～2013 年 07 月 18 日, 福岡県福岡市

⑤ 村井 光, 吉村 友太, 岡本 康宏, 正代 隆義, 宮原 哲浩, 辺縮約に基づく木構造パターンの列挙とそのデータマイニングへの応用, 火の国シンポジウム 2013, 情報処理学会九州支部, 2013 年 03 月 14 日～2013 年 03 月 15 日, 熊本大学工学部

⑥ 岡本 敦, 正代 隆義, ストリーム上の頻出時系列とその近似発見アルゴリズムについて, 情報処理学会第 75 回全国大会, 2013 年 03 月 06 日～2013 年 03 月 08 日, 東北大学川内キャンパス

⑦ 岡本 康宏, 吉村 友太, 正代 隆義, 辺縮約に基づく木構造パターンの多項式時間学習アルゴリズム, 電気関係学会九州支部第 65 回連合大会, 2012 年 09 月 24 日～2012 年 09 月 25 日, 長崎大学工学部

⑧ T. Hino, Y. Suzuki, T. Uchida, and Y. Itokawa, Polynomial Time Pattern Matching Algorithm for Ordered Graph Patterns, The 22nd Int. Conf. Inductive Logic Programming (ILP2012), 2012 年 09 月 17 日～2012 年 09 月 19 日, Dubrovnik, Croatia

⑨ Y. Yoshimura and T. Shoudai, Unordered Tree Contraction Patterns in Polynomial Time, The 22nd Int. Conf. Inductive Logic Programming (ILP2012), 2012 年 09 月 17 日～2012 年 09 月 19 日, Dubrovnik, Croatia

⑩ 岡本 敦, 鶴田 悠, 正代 隆義, 頻出時系列発見近似ストリームアルゴリズムとそのデータスクリーニングへの応用について, 火の国シンポジウム 2012, 情報処理学会九州支部, 2012 年 3 月 15 日, 九州工業大学情報工学部

⑪ T. Yamada and T. Shoudai, Graph Contraction Pattern Matching for Graphs of Bounded Treewidth, The 21st Int. Conf. Inductive Logic Programming (ILP 2011), 3rd August, 2011, Windsor Great Park, United Kingdom

⑫ Y. Yoshimura, T. Shoudai, Y. Suzuki, T. Uchida, and T. Miyahara, Polynomial Time Inductive Inference of Cograph Pattern Languages from Positive Data, The 21st Int. Conf. Inductive Logic Programming (ILP 2011), 3rd August, 2011, Windsor Great Park, United Kingdom

６．研究組織

(1)研究代表者
正代　隆義（SHOUDAI, Takayoshi）
九州大学・大学院システム情報科学研究院・准教授
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

(3)連携研究者
溝口　佳寛（MIZOGUCHI, Yoshihiro）
九州大学・マスフォアインダストリ研究所・准教授
研究者番号：80209783