様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月２２日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19700145
研究課題名（和文）　領域知識を制約として用いるグラフマイニング手法の実現

研究課題名（英文）　Development of Graph Mining Method Using Domain Knowledge as Constraints

研究代表者
大原　剛三（OHARA KOUZOU）
大阪大学・産業科学研究所・助教
研究者番号：30294127

研究成果の概要：本研究では，化合物の構造式のようなグラフ構造データを大量に蓄積したデータベースから特徴的な部分パターンを発見するグラフマイニングにおいて，領域知識に基づいた構造的制約を導入することにより，無駄な部分パターン候補を合理的，かつ効果的に削減する手法を実現した．肝炎患者の検査履歴データに提案手法を適用した評価実験においては，従来手法では発見できなかった，肝炎の進行状況推定に有用な部分構造を現実的な実行時間内で発見できることを確認した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,700,000	0	1,700,000
２００８年度	1,600,000	480,000	2,080,000
年度			
年度			
年度			
総　計	3,300,000	480,000	3,780,000

研究分野：知識発見，知識工学
科研費の分科・細目：情報学・知能情報学
キーワード：データマイニング，グラフマイニング，機械学習，領域知識

１．研究開始当初の背景

コンピュータの急速な普及に伴い様々な分野において膨大な電子化情報が蓄積されつつある今日，その表現形式も従来の関係データベースに代表される単純な表形式から，化学物質の分子構造のような構造データ，HTML のような木構造データ，自然言語のような半構造データなど多岐にわたり，そのような構造をもつデータから有用な知識を発掘するデータマイニング技術の開発が急務となっている．これらのデータは頂点，および頂点を結ぶ辺により構成されるグラフとして表現可能であるため，近年，グラフ構造から有意な塊（部分パターン）を発掘するグラフマイニングが盛んに研究されるようになった．これまでに AGM，GBI などの先駆的なグラフマイニングアルゴリズムをはじめ，FSG，gSpan など多数のグラフマイニングアルゴリズムが提案されている．これらのアルゴリズムは，所与のグラフ集合から高い頻度で現れる多頻度部分グラフをいかに効率的に抽出するかに主眼を置いたものである．しかしながら，多頻度パターンは実際の応用，すなわち特定の対象領域における特定の目的下では必ずしも有用なものとは限らない．なぜなら，データ中に多頻度で現れることは，

対象領域の専門家にとっては既知のパターン・傾向である可能性が極めて高いことを意味するからである．実際，対象領域の専門家からは「興味深い傾向を含むパターンの探索」や「既知の傾向を含まないパターンの発見」などに対する要求が極めて高いが，そのような領域知識に基づいたパターン発掘が可能なグラフマイニングアルゴリズムは国内外を問わず皆無であり，また既存の頻度指向のグラフマイニングアルゴリズムを利用しても膨大な多頻度パターンが計算機資源を圧迫するため真に必要なパターンの発見に至らないのが現状である．

２．研究の目的

以上のような背景から，本研究では，領域知識をグラフマイニングのパターン発見過程において制約として利用することで膨大な探索空間を合理的かつ効果的に制限し，領域専門家にとってより興味深い，もしくは対象問題を解決する為により有用なパターンを高速に発見できる手法を実現することを目的とする．そのために，以下の部分目標を設定する．

（１）領域知識に基づく制約をパターン発見過程において探索空間の削減に用いる手法の実現．

領域知識に基づく制約として，部分パターンとして表現可能な構造的制約を考え，(A) 領域専門家が興味をもつ傾向を示す部分パターン，および (B) 領域専門家にとって既知の傾向を示し興味の対象にはなり得ない部分パターンを制約パターンとし，パターン発見過程において探索空間中の候補パターンを効率的に削減する手法の実現を目指す．

（２）制約充足判定にかかるコスト，およびパターン探索アルゴリズム自体のコストを削減する要素技術の開発．

領域知識に基づく制約を導入した場合，制約の充足性判定にかかるコストが新たに発生するため，制約を満たすパターンの高速発見のためには制約充足判定にかかるコスト，およびパターン探索アルゴリズム自体のコスト削減が必須となる．そのため，以下の技術目標の達成を目指す．

①部分パターンの評価値に基づく効果的な枝刈り手法の開発．

②グラフを表現する行列の固有値を用いた部分グラフ同型性判定の効率化．

３．研究の方法

前述の各研究目標・技術課題に対する本研究におけるアプローチについて述べる．

（１）領域知識に基づく制約をパターン発見過程において探索空間の削減に用いる手法の実現について．

前述の (A)，(B) 2 種類の制約パターンをパターン発見過程において対象データと統一的に扱う枠組みとして，本研究では研究代表者がこれまでに開発に取り組んできたグラフマイニングアルゴリズム Cl-GBI を提案手法の基礎とする．Cl-GBI は，グラフ中の隣接する 2 つの頂点を逐次的に統合し擬似頂点を生成することを再帰的に繰り返すものであり，結果として得られた擬似頂点が部分パターンを表す．Cl-GBI のアルゴリズムを以下に示す．なお，擬似頂点を生成する操作を擬似チャンキングと呼ぶ．

［入力］
G ：グラフデータベース
b ：一度に生成される擬似頂点数
L ：最大繰り返し回数
C ：擬似チャンキングされる頂点ペアの順位基準
θ ：抽出される部分構造が満たすべき条件
［出力］
S ：θ を満足する部分パターン集合

Step1．G 中の隣接するすべての頂点ペアを抽出する（2 回目以降の繰り返しでは，少なくとも 1 つの頂点が擬似頂点であるような頂点ペアを抽出する）．
Step2．抽出した頂点ペアの頻度を計算し，条件 θ を満たさないものを削除する．
Step3. 順位基準 C の下で上位 b 個の頂点ペアを選択し，新しい擬似頂点として S に登録する（2 回目以降の繰り返しでは，それまでに選択されなかった頂点ペア，および新たに抽出された頂点ペアの中から上位 b 個を選択する）．
Step4. 新しい擬似頂点に新しいラベルを割り当て，繰り返し回数が L 回未満であれば Step1 に戻る．

ここでは，上記アルゴリズムの Step1 において，頂点ペアが制約パターンを含むか否かを判定することで，Cl-GBI の探索空間を構造的制約で制限することを考える．

（２）制約充足判定にかかるコスト，およびパターン探索アルゴリズム自体のコストを削減する要素技術の開発について．

①部分パターンの評価値に基づく効果的な枝刈り手法の開発について．

ここでは，特徴的な部分パターンの評価基準として情報利得を考える．クラスが付与されたデータに対して，情報利得が高い部分パターンほどクラス分類への寄与が高い，すなわちいずれかのクラスの特徴をよく反映していると解釈できる．パターン探索アルゴリズムでは，これまでに得られている最大の情報利得値よりも大きな情報利得値を得ることが期待できない部分パターンは早期に探索

空間から排除されるべきであり，ここではその為の判定基準と，その判定基準を盛り込んだアルゴリズムの実現を図る．具体的には，情報利得が凸関数であることから，当該判定基準の実現には凸関数のもつ性質を利用する．

②グラフを表現する行列の固有値を用いた部分グラフ同型性判定の効率化について．

ここでは，グラフが隣接行列などの行列で表現可能であり，あるグラフとその誘導部分グラフの隣接行列の固有値の間には Interlace 定理と呼ばれる関係が存在することに着目する．隣接行列とは，頂点数 n 個のグラフを $n \times n$ 正方行列で表現したものであり，行列の (i, j) 要素はグラフ中の i 番目の頂点と j 番目の頂点間の辺の有無を 0 か 1 で表す．グラフ g' があるグラフ g の誘導部分グラフである場合，g' の隣接行列は g の隣接行列の主小行列となる．一方，Interlace 定理は，$n \times n$ 行列の固有値 $\lambda_1, \cdots, \lambda_n \ (\lambda_1 \le \cdots \le \lambda_n)$ とその $m \times m$ 主小行列 $(m \le n)$ の固有値 $\lambda'_1, \cdots, \lambda'_m \ (\lambda'_1 \le \cdots \le \lambda'_m)$ に以下の関係が成り立つことを示すものである．

$$\lambda_k \le \lambda'_k \le \lambda_{k+(m-n)} \quad (k = 1, \cdots, m)$$

本研究で導入する構造的制約の充足判定には，あるグラフが他のグラフの部分グラフとなっているか否かを判定する部分グラフ同型性判定が必須となるが，部分グラフ同型性判定の計算複雑さは NP 完全であることが知られている．その為，ここではより計算量の少ない隣接行列の固有値を計算し，Interlace 定理を利用することで，厳密な部分グラフ同型性判定を可能な限り回避することを考える．

４．研究成果

（１）領域知識に基づく制約をパターン発見過程において探索空間の削減に用いる手法の実現に関する成果

構造的制約の充足判定には，計算複雑さが NP 完全である部分グラフ同型性判定が必須となる為，ここでは頂点ペアに関してグラフの構造や頂点・辺のラベルからグラフごとに一意に定まるグラフ不変量を擬似頂点生成過程において逐次的，かつ効率的に計算・更新し，その値を制約パターンのグラフ不変量と比較することで，部分グラフ同型性判定を実施する必要があるか否かを事前に判断する構造的制約充足判定手法を実現した．具体的には，辺の接続関係を考慮するとグラフ不変量の計算自体に時間がかかるため，個々の頂点と辺のラベルの出現回数のみに基づいたグラフ不変量だけを利用した．

提案手法を実データである肝炎患者データベースに適用した評価実験の結果を表１に示す．本実験では，当該データの従来の解析結果から得られる知見に基づき，4 種類のパターンを結果として得られる特徴的な部分パターンに含まれるべき制約パターンとして与えた．その結果，表１に示すように制約を用いない場合よりも短い実行時間で，同程度かそれ以上の情報利得値をもつ，すなわち肝炎患者の病状進行レベル分類により寄与する部分パターンの発見に成功した．

表１ 制約パターンを用いた評価実験の結果

用いた制約パターン	実行時間（秒）	情報利得の最大値
なし	43,971	0.1139
No.1	9,344	0.1076
No.2	6,720	0.1698
No.3	20,383	0.1110
No.4	4,913	0.1297

領域知識に基づく部分パターンをグラフマイニングにおけるパターン探索における制約として直接利用できるグラフマイニング手法は極めて新規性が高く，今後，様々な応用分野で，領域専門家にとってより興味深い部分パターン（知見）の発見に大きく寄与するものと考えられる．

（２）制約充足判定にかかるコスト，およびパターン探索アルゴリズム自体のコストを削減する要素技術の開発に関する成果

①部分パターンの評価値に基づく効果的な枝刈り手法の開発に関する成果

一般に，凸関数に関しては関数値の上界を求めることが可能であり，情報利得もまた凸関数である．この上界が計算可能であるという性質を利用し，それまでに得られた情報利得の最大値よりも大きな情報利得を得ることが期待できる部分パターンを生じさせない頂点ペアを Cl-GBI の探索空間から削除する枝刈り手法を実現した．具体的には，頂点ペアにより表される部分パターン g のクラス別の頻度がわかっている場合に，g を包含するパターンにより得られる情報利得の上界 u，および g のクラス別頻度ではなく，g を構成する各頂点のクラス別頻度がわかっている場合に g を包含するパターンにより得られる情報利得の上界の見積値 $\hat{u}$ を計算し，それらが現在の情報利得値の最大値 τ 以下の場合は，g を探索空間から削除する．ただし，$u \le \hat{u}$ であり，$\hat{u}$ は g の各頂点を含むグラフ集合の積集合を g を含むグラフ集合とみなし g のクラス別頻度を計算することにより得られる．

u は，Cl-GBI のアルゴリズム Step2 で g の頻度を計算した後で計算可能であることから，u を用いた枝刈りを事後枝刈りと呼ぶ．これに対し，$\hat{u}$ は Cl-GBI のアルゴリズム

Step2 での g の頻度計算の前に計算可能であることから，$\hat{u}$ を用いた枝刈りを事前枝刈りと呼ぶ．

表 2 に，事前枝刈り・事後枝刈りを導入した Cl-GBI を肝炎患者データベースに適用した評価実験の結果を示す．Cl-GBI の繰り返し回数を 15 とした場合には，枝刈りをしない場合と比べて両枝刈りを適用すると約 80%の実行時間の削減を実現しており，枝刈りの効果が非常に高いことが確認できた．

表 2　情報利得値に基づく枝刈りを導入した Cl-GBI の評価実験結果

枝刈り条件	実行時間（秒）	
	L=10	L=15
なし	534	1,894
事前枝刈り	380	1,430
事後枝刈り	140	362
事前・事後枝刈り	137	353

なお，事前枝刈りでは，頂点ペアの頻度の見積値が実際の頻度よりもかなり高い目になる傾向がある為，枝刈り条件を満足する頂点ペアの数が少なくなり，その効果は限定的である．しかしながら，頂点の隣接関係などグラフの構造的特徴を利用することで，より厳密な頻度の見積りが可能であることから，さらなる効果の向上が期待できる．

②グラフを表現する行列の固有値を用いた部分グラフ同型性判定の効率化に関する成果

Interlace 定理では，2 つのグラフが誘導部分グラフ同型となる必要条件を示すだけであり，Interlace 定理による固有値の条件を満たしていても，誘導部分グラフ同型とならない場合もある．実際のところ，単純に Interlace 定理を適用しただけではその判定精度が低いため，本研究ではグラフの隣接行列の要素の値（当該行列の固有値）を Interlace 定理による誘導部分グラフ同型性判定の精度を高めるように最適化する手法を提案した．具体的には，隣接行列の要素の値を対応する頂点と辺により一意に定まるグラフ特徴に割り当てられた実数と解釈し，そのグラフ特徴と実数の対応関係を定める関数 F に対し，以下のような $|D(g',F)|$ を定義し，その値を最小にするように F を最適化する手法を提案した．

$$|D(g',F)| = \sum_{i=1}^{|G|} \sigma(g_i, g', F)$$

ただし，

$$\sigma(g_i, g', F) = \begin{cases} 1 & g' \subseteq_{IN} g_i \\ 0 & otherwise \end{cases}$$

とし，$g' \subseteq_{IN} g_i$ は g' が g の誘導部分グラフであることを意味する．また，$g_i \in G\,(i = 1, \cdots, |G|)$ である．なお，F の最適化には最急降下法を用いた．

人工データを用いた評価実験では，30 個のグラフに対して，それらの誘導部分グラフではない g' との誘導部分グラフ同型性を Interlace 定理で判定した場合，関数 F の最適化前では 30 個すべてのグラフについて g' を誘導部分グラフとなる可能性があると判定したのに対し，最適化後では 30 個すべてのグラフついて g' を誘導部分グラフとなる可能性はないと判定しており，最適化による判定精度の大幅な向上を確認した．実際には，Interlace 定理により誘導部分グラフ同型である可能性があると判断されたグラフの組に対してのみ厳密な誘導部分グラフ同型性判定を実施すればよい．

（誘導）部分グラフ同型性判定に Interlace 定理を用いる試みはこれまでもあったが，判定精度を改善する為に行列要素の最適化を試みた事例はなく，提案手法は極めて新規性の高いアプローチである．提案手法は，グラフデータベース G と特定のグラフ g' を入力とするものであるが，G を 1 つのグラフからなるものと考えれば，1 対 1 の誘導部分グラフ同型性判定の判定精度を高めることが可能である．今後，誘導部分グラフに限らず，一般部分グラフへも拡張することで，グラフマイニングに限らず，グラフに関する多くの問題の解決に寄与することが期待できる．

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１件）

① Alexandre Termier, Marie-Christine Rousset, Michele Sebag, Kouzou Ohara, Takashi Washio, and Hiroshi Motoda, DryadeParent, an Efficient and Robust Closed Attribute Tree Mining Algorithm, IEEE transactions on Knowledge and Data Engineering, Vol.20, No.3, pp.300-320 (2008), 査読有り．

〔学会発表〕（計９件）

① グェン ズィ ヴィン, 大原剛三, 鷲尾隆, 部分グラフ同型性判定のためのグラフスペクトル最適化手法, 第 9 回人工知能学会データマイニングと統計数理研究会（SIG-DMSM), 2009 年 3 月 4 日, 京都・メルパルク京都.

② Kouzou Ohara, Masahiro Hara, Kiyoto Takabayashi, Hiroshi Motoda, and Takashi Washio, Pruning Strategies based on the Upper Bound of Information

Gain for Discriminative Subgraph Mining, The 2008 Pacific Rim Knowledge Acquisition Workshop (PKAW2008), 2008年12月16日, ベトナム・ハノイ工科大学.

③ Kouzou Ohara and Takashi Washio, Isomorphism Identification by Using Graph Spectra and Its Application to Graph Mining, Joint Meeting of 4th World Conf. of the IASC and 6th Conf. of the Asia Regional Section of the IASC on Computational Statistics & Data Analysis (IASC2008), 2008年12月6日, 神奈川・パシフィコ横浜.

④ 大原剛三, 鷲尾隆, 大規模グラフにおける近似部分グラフ検索手法, 2007年度人工知能学会全国大会,2007年6月22日, 宮崎・ワールドコンベンションセンターサミット.

⑤ 高林健登, 原昌弘, 大原剛三, 元田浩, 鷲尾隆：情報利得値の上界に着目した特徴的部分グラフの効率的なマイニング, 2007年度人工知能学会全国大会, 2007年6月22日, 宮崎・ワールドコンベンションセンターサミット.

６．研究組織

(1)研究代表者

大原　剛三 (OHARA KOUZOU)

大阪大学・産業科学研究所・助教

研究者番号：30294127

(2)研究分担者

(3)連携研究者