3F2-01

部分グラフを制約とするグラフ構造データからの知識獲得

Knowledge Acquisition from Graph Structured Data with Constraints

茂木 明 Akira Mogi　吉田 哲也 Tetsuya Yoshida　ワロドム ジアムサクン Warodom Geamsakul　大原 剛三 Kouzou Ohara　元田 浩 Hiroshi Motoda　鷲尾 隆 Takashi Washio

大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Osaka University

A machine learning technique called Beam-wise Graph-Based Induction (B-GBI) can extract discriminative patterns from graph structured data by stepwise pair expansion (pairwise chunking). The extracted patterns are useful as classifier, but not always useful for domain experts because in most cases they are trivial or difficult to understand for experts. To improve B-GBI so as to acquire more useful patterns for experts, we think using domain knowledge they possess is a useful and reasonable approach. As an instance of such improvement, in this paper, we propose a method of refining domain knowledge by using a subgraph related to the knowledge as a constraint to limit patterns to be extracted. In this method, we assume a binary classification setting and that domain knowledge is not always correct, i.e., there are some exceptions to it. The proposed method divides a given dataset into two subsets: one is a set of graphs containing the restrictive pattern, and the other is a set of remaing graphs. Then it extracts patterns that are contained in graphs belonging to a class, but not in those belonging to the other class, from the former set by means of B-GBI. The extracted patterns can refine the given domain knowledge in a sense that they exclude exceptions. Furthermore we show the results of preliminary experiments using a hepatitis dataset and discuss the usefulness of patterns acquired by the proposed method.

1. はじめに

データからの知識発見では複雑な構造をもつデータを対象とする場合が多数存在し，そのようなデータの表現方法としてはグラフ表現が有用である．たとえば，化学物質の分子構造などは属性と値のペアを表形式で記述する従来のデータ表現方法で表現することは困難であるが，グラフ構造を用いれば自然に表現することができる．このような背景から，グラフ構造データからの知識発見は重要な研究課題であるといえる．

有向グラフから類型パターンを発見する一つの手法として Graph-Based Induction（GBI 法）がある [吉田 97, 松田 01]. GBI 法では，グラフ中のノードのペアを逐次拡張していくことを基本原理とし，ペアを選択するための評価関数としては主に統計的指標を用いている．このような GBI 法に対して，松田らはクラス分類能力のあるパターンを発見するために，2 種類の評価関数，ビーム探索，および Canonical Label [Fortin 96] の考えを GBI 法に取り入れた Beam-wise Graph-Based Induction（B-GBI 法）を提案した [松田 02].

従来の B-GBI 法では，与えられた評価関数の閾値を満たすパターンを機械的に近似探索するため，結果的に得られるパターンが対象領域の専門家にとっては既知の内容であったり，また理解が困難なものであったりするなど，必ずしも専門家にとって有用であるとは限らなかった．このような問題に対して，筆者らは専門家が既にもっている領域知識を B-GBI 法の探索に反映させることが，より有用なパターン（知識）を獲得する上で有効であると考えた．本稿ではそのような考えに基づき，領域知識に基づいたパターン（部分グラフ）を制約として用いる知識の精緻化手法を提案する．提案手法では，制約として与えられたパターンが含まれるグラフをまず取り出し，その集合において B-GBI 法を適用することで，領域知識を精緻化するために有用なパターンを獲得する．また，本稿では肝炎患者に対するインターフェロン薬剤の奏功性に関するデータに対して提案手法を適用した予備実験の結果を示し，得られたパターンの有用性について議論する．

連絡先: 茂木　明
〒 567-0047　大阪府茨木市美穂ヶ丘 8-1
大阪大学産業科学研究所 元田研究室
電子メール：mogi@ar.sanken.osaka-u.ac.jp

2. Beam-wise Graph-Based Induction（B-GBI 法）

2.1 GBI 法と B-GBI 法

GBI 法はグラフ構造データ中に現れる特徴的なパターンを抽出することを目的に考案された [吉田 97]．GBI 法は図 1 に示すように「ペアの逐次抽出（チャンキング）により特徴的なパターンを抽出する」という基本的な考えにより実現されている．ここで,「ペア」とは「二つのノードおよびそれらをつなぐリンク」からなる GBI 法で用いる基本単位となるものである．また，ペアは逐次拡張される（チャンクされる）ことにより複雑なパターンになっていく．

GBI 法では，評価値が同値のペアが複数存在した場合や，同じノードラベルが多数存在するために同種類のペアの連鎖が生じた場合には，チャンクすべきペアの選択に曖昧性が生じる．しかしながら，一度チャンクしたペアを元の個別のノードに復元して評価しないため，その探索は Greedy 探索となる．

以上のような GBI 法に対し，B-GBI 法はクラス分類能力のあるパターンを「特徴的なパターン」として抽出するために，以下の 3 点について GBI 法を改良したものである．

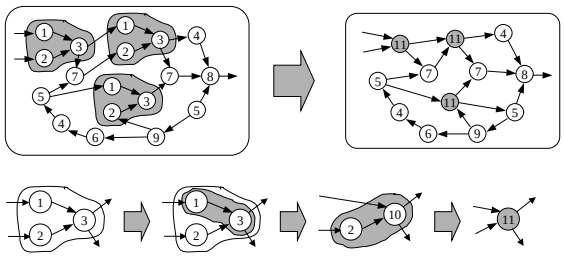


図 1: 逐次ペア拡張の基本的な考え

- 2 種類の評価指標の導入
  GBI 法では頻度の高いペアをチャンクすることでパターンを抽出するが，その結果得られたパターンが必ずしもクラス分類能力が高いとは限らない．一方，情報利得 [Quinlan 86] などのクラス分類能力に関する指標をペアの選択に用いた場合，パターンのサイズに対してその評価は非単調に変化するため，あるペアの評価が高かったとしても，そのペアが将来的にクラス分類能力の高いパターンに成長する保証はない．そのため，B-GBI 法では頻度を基準にペアを選択することである程度意味のある大きなパターンを生成しつつ，その中から別の評価関数を用いてクラス分類能力の高いパターンを抽出している．

- ビーム探索法の導入
  GBI 法では，Greedy 探索を用いることで，ある程度の大きさを持った有意な部分グラフを現実的な時間で抽出することが可能となる反面，得られた「特徴的なパターン」が「もっとも特徴的なパターン」である保証はない．この問題を低減するために，B-GBI 法ではビーム探索を導入している．具体的には，各時点でチャンクするペアをビーム幅分だけ選択し，常にビーム幅分の状態を並列に維持することで，状態数の爆発を回避しつつ，探索空間の拡張を実現している．

- Canonical Label の導入
  GBI 法では，パターンをチャンクして 1 つのノードに書き換えるため，ペアを数え上げるときに同じグラフ構造にもかかわらず，異なったペアとして扱ってしまう場合がある．これを防ぐためにグラフのノード・リンクの要素を含めた構造を Canonical Label で表現し，要素のラベルが同じものを同じグラフ構造と判定している．

2.2 B-GBI 法のアルゴリズム

B-GBI 法が行う「逐次ペア拡張」アルゴリズムは以下の通りである．この作業は与えられた終了条件が満たされるまで繰り返される．

ステップ 1 全ての状態について，グラフに存在するペアを全て抽出する．

ステップ 2a ステップ 1 で抽出したペアのうち，評価関数に基づき特徴的なペアを全て登録する．この時，ペアを構成するノードが既に書き換えられたノードであれば元のパターンに復元してから登録する．

ステップ 2b ステップ 1 で抽出したペアのうち，頻度によりチャンクすべきペアをある一定の数だけ選び，抽出パターンとして登録する．このときペアを構成するノードが既に書き換えられたノードであれば元のパターンに復元してから登録する．もしチャンクすべきペアがなくなれば終了する．

ステップ 3 ステップ 2b で選ばれたそれぞれのペアに対し，ペアを一つのノードに置き換えることで，それぞれのグラフを書き換える．この際，必要に応じて状態を分裂または消滅させる．そして，ステップ 1 に戻る．

3. 制約を用いた B-GBI 法による知識獲得

3.1 制約の必要性

従来の B-GBI 法においては，専門家にとって理解容易でないパターンが出力されることや，既知のパターンが出力されることがあった．そのような問題を回避するためには，出力されるパターンの形式や種類，もしくはパターンの探索過程を制御する何らかの制約を導入することが有用であると考えられる．そのような制約の導入に関しては様々な方法が考えられるが，本稿では専門家のもつ領域知識をパターンの探索における制約として用いることを考える．専門家のもつ領域知識を制約として事前に与えることで，専門家にとって既知のパターンが出力されるこを回避できるとともに，そのような制約下で成立する有用なパターンが新たに獲得できることが期待できる．以下では，制約下で成立する有用なパターンとして，制約として与えられた領域知識をより精緻化する効果をもつパターンを B-GBI 法を用いて獲得する手法を提案する．

3.2 領域知識の精緻化手法

ここでは，グラフ g が“ パターン A を含むならば概ねクラス Y である ”というタイプの領域知識を仮定する．また，対象とするクラスは 2 クラスであるとする．このとき，この領域知識を精緻化することは“ 概ね ”というあいまいさを排除することを意味する．そのためには，パターン A を含むグラフに関して，真にクラス Y であるグラフ間，もしくは真にクラス Y でないグラフ間に特徴的な（1 つ以上の）パターンを獲得できればよい．そうすれば，例えば“ パターン A を含み，かつパターン B，C を含むなら必ずクラス Y である ”などのような確固たる知識を得ることができる．以下，B，C などのようなパターンを精緻化パターンと呼ぶ．

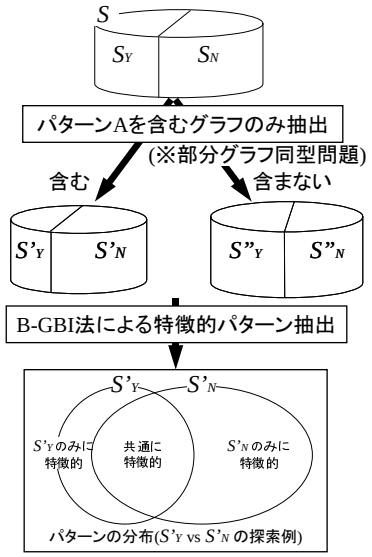


図 2: 領域知識の精緻化

手法の概要を図 2 に示す．提案手法では，まず入力として与えられた事例集合（グラフ集合）S を，領域知識として利用されるパターン A を含む集合と含まない集合の 2 つに分割する．利用する領域知識が完全に正しいわけではないことから，結果的に各集合は，2 つのクラス Y，および N それぞれの事例を含むことになる．その後，パターン A を含むグラフ集合に着目し，B-GBI 法を用いて，当該集合中の各クラスに属するグラフの集合 S'_Y，S'_N 各々に関して特徴的なパターンを精緻化パターンとして抽出する．ここで，S'_Y に関して特徴的なパターンとは，S'_Y 中のグラフには含まれるが S'_N 中のグラフには含まれず，かつ頻度の高いものとする．S'_N に特徴的なパターンも同様である．

このような精緻化のメリットとしては，得られる知識が直観

的に理解しやすいという点が挙げられる．たとえば，決定木では各分岐における分離テストは直前の分岐における分離テストを精緻化していると捉えることができるが，その解釈は“ 属性 X の値が a でなければクラス N である ”などのように,「〜でなければ」という否定的な条件になることが多分にある．しかしながら，実世界におけるデータでは，通常，属性 X の値が何かしらの値であることを陽に記述するが，ある値ではないことは陽には記述しないことから，このような否定的な条件は直観的に理解しづらいものといえる．これに対し，ここでの精緻化手法では「得られたパターンが含まれるならば」という形式の肯定的な条件に必ずなることから，得られる知識は比較的理解しやすいものとなる．

3.3 部分グラフ同型問題の近似解法

前節で述べた手法を実現するためには，それぞれのグラフがパターン A を含むか否かを判定する必要があり，それは部分グラフの同型問題を解くということに等しい．しかしその計算量は非常に膨大であるため，ここでは B-GBI 法を利用した部分グラフ同型問題の近似解法を提案する．

基本的な考えは，B-GBI 法を用いることにより，対象とするグラフからパターン A が抽出できるかどうかを現実的な時間内で判定するというものである．ペアの逐次拡張により得られるパターンは必ずそのグラフ中に存在することから，この手法でパターン A が含まれると判定された場合，それは必ず正しい．その反面，B-GBI 法もまたその探索は Greedy 探索であるため，グラフ中のパターン A を必ず見つけ出せる保証はない．すなわち，この手法は健全ではあるが完全ではない．以下にその手順をまとめる．

1. 判定したいパターンに含まれるリンク・ノードラベル以外の要素をグラフからすべて取り除く．
2. B-GBI 法による探索を行い，特徴的パターンを抽出する．
3. 抽出された部分グラフの Canonical Label と，判定したいパターンの Canonical Label とを計算し，照合する．
4. Canonical Label が一致した場合，そのグラフ番号を出力する．

この手法では，探索するグラフから含めたい部分グラフの要素以外を取り除くことで必要な要素のみを残し，探索空間を小さくすることによって，部分グラフを発見しやすくしている．

4. 予備実験

4.1 実験設定

提案手法により得られる精緻化パターンの性質を調べるために，実データを対象にした予備実験を行った．対象データは，千葉大学医学部付属病院からご提供いただいた慢性肝炎患者に対するインターフェロン薬剤の効果の有無に関するデータセット [山口 02] であり，事例数（患者数）は 94 である．各事例は，各患者に対する一定期間に複数回実施された各種検査の検査値からなり，インターフェロン薬剤の効果の有無に応じて以下の 2 クラスに分類される．

クラス R: 投与の効果があった患者（患者数:38）

クラス N: 投与の効果がなかった患者（患者数:56）

実験では，各検査項目をリンク，2 週間単位で平均化した検査値をノードとして 1 患者分のデータを 1 つの連結グラフで表現した．また，精緻化する元の領域知識で用いるパターンとしては，2 つのクラスの事例に共通に現れるパターンのうち最も情報利得が高いものを選んだ．そのパターンを図 4.2 に示す．

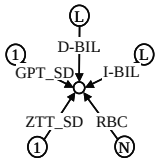


図 3: パターン A

このパターン A を含むグラフを 3.3 節で述べた部分グラフ同型問題の近似解法で求めた結果，94 個の全グラフ中 48 個のグラフがパターン A を含み，そのうち 38 個のグラフがクラス N であった．このことから，本実験では“ パターン A を含むなら概ねクラス N である ”という知識を精緻化することになる．以下，パターン A を含み，かつクラスが R であるグラフの集合を S'_R，パターン A を含み，かつクラスが N であるグラフの集合を S'_N とする．

一方，前述のように本稿における精緻化では，S'_R に含まれ他方には含まれないパターンを精緻化パターンとして獲得するが，その際，パターンが含まれないことを確認する集合としては，S'_N の他に，クラス N 全体の集合である S_N，および S'_R 以外の全事例集合 $\overline{S'_R}$ などを利用することができ，それぞれの場合で，得られる精緻化パターンの意味合いが異なる．たとえば，S_N を対象とした場合，得られる精緻化パターンはクラス N のグラフには絶対現れないクラス R に関して特徴的なパターンであるといえるが，S'_N を用いた場合には，S'_N に含まれないクラス N のグラフにはそのパターンが現れる可能性があることから，当該パターンはあくまで元のパターン A を前提とした上でクラス R に関して特徴的であるといえる．S'_N から精緻化パターンを獲得する場合も，同様の議論が成り立つ．

本実験では，このようにパターンが含まれないことを確認する集合の違いによって生じる精緻化パターンの違いを調べるために，以下の 6 通りの精緻化を行い，それぞれの場合に関して，特徴的なパターンを求めた集合に含まれるグラフを可能な限り多く被覆するために必要な精緻化パターン数を，情報利得の高いものから順に上位 100 個までを対象として調べた．なおここでは，あるパターン P がグラフ g を被覆するとは，g が P を部分グラフとして含むことを意味する．また，本実験では，全ての B-GBI 法に関して，そのビーム幅を 3 とした．

Case1: S'_R に特徴的なパターンを S'_N を利用して獲得
Case2: S'_R に特徴的なパターンを S_N を利用して獲得
Case3: S'_R に特徴的なパターンを $\overline{S'_R}$ を利用して獲得
Case4: S'_N に特徴的なパターンを S'_R を利用して獲得
Case5: S'_N に特徴的なパターンを S_R を利用して獲得
Case6: S'_N に特徴的なパターンを $\overline{S'_N}$ を利用して獲得

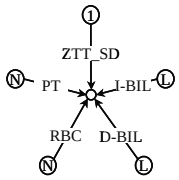
4.2 結果と考察

実験結果を 1 に要約する．表 1 における Set1 は特徴的なパターンを求めた集合であり，Set2 はその際にパターンが含まれないことを確認した集合である．また，All とは Set1 に含まれるすべてのグラフを可能な限り被覆するために必要な精緻化パターンの数とそれらの精緻化パターンにより実際に被覆されるグラフ数を示しており，Majority とは Set1 に含まれるグラフの過半数を対象にした場合である．なお，精緻化パターンの選択に関しては，情報利得の高いものから順に，それまでに選択されたパターンによって被覆されていないグラフを新たに被覆するようなパターンを選択した．

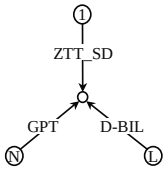
表 1: 得られたパターンの数とそれにより被覆されるグラフの数

Case	Set1	Set2	All	Majority
1	S'_R	S'_N	3 パターン (グラフ数:6/10)	3 パターン (グラフ数:6/10)
2	S'_R	S_N	7 パターン (グラフ数:10/10)	4 パターン (グラフ数:6/10)
3	S'_R	$\overline{S'_R}$	6 パターン (グラフ数:9/10)	3 パターン (グラフ数:6/10)
4	S'_N	S'_R	9 パターン (グラフ数:29/31)	3 パターン (グラフ数:19/31)
5	S'_N	S_R	9 パターン (グラフ数:24/31)	5 パターン (グラフ数:16/31)
6	S'_N	$\overline{S'_N}$	10 パターン (グラフ数:22/31)	6 パターン (グラフ数:16/31)

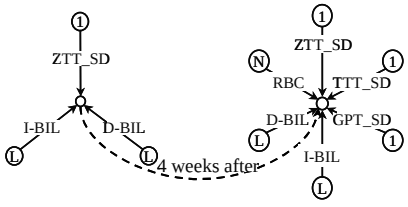
表 1 を見ると，場合によってばらつきがあるものの，複数のパターンによって Set1 中のグラフの大多数を被覆できていることがわかる．このことから，提案手法により獲得したパターンは知識の精緻化に有用であるといえる．実際に獲得された精緻化パターンの例として，Case4 の Majority における 3 つのパターンを図に図示する．これら 3 つのパターンを用いることで，“ パターン A を含み，かつパターン 1，パターン 3，パターン 9 のいずれかを含むグラフは**必ず**クラス N である ” という，精緻化された知識を得ることができる．ここで，図における 3 つのパターンと図におけるパターン A を比較すると，図における各パターンはパターン A と多くのリンクラベルとノードラベルの組合せを共有しており，一部だけが異なっていることがわかる．この異なった部分が精緻化において意味をもつものと思われる．また，パターン 9 は，時間的変化を含んだパターンとなっており興味深い．今後，これら精緻化パターンの解釈・評価を専門家から受けることが重要となる．



(a) パターン 1



(b) パターン 3



(c) パターン 9

図 4: Case4 実行時の情報利得の上位 3 パターン

一方，表 1 において Case1〜3，および Case4〜6 をそれぞれ下に見ていくと，Set1 中のグラフを被覆するために必要なパターン数が増加する傾向にあることがわかる．言い換えるなら，1 つの精緻化パターンが被覆する Set1 のグラフ数が減少する傾向にあるといえる．これは，Case1 に対して Case2, Case3 ほど，Set2 のグラフ数が多くなることから，Set1 から得られる精緻化パターンの候補，すなわち Set1 中のグラフを被覆し，かつ Set2 中のグラフは 1 つも被覆しないパターンの数が減少するためだと考えられる．Set1 中のグラフをより多く被覆するパターンは，元々パターン A を含むという点で Set1 中のグラフと類似している Set2 中のグラフも被覆する可能性が高い．このことから，より少ないパターンで多くのグラフを説明するためには，Set2 を相対的に小さくすればよいが，それは得られるパターンが Set1 のみを特徴づけるという観点からすれば，必ずしもよいパターンを見つけることにはならない．Set2 を相対的に小さくした場合，得られるパターンは Set2 に含まれない他の多数のグラフにも含まれ得る．逆に，Set2 を相対的に大きくすることは，少ないパターンで多くのグラフを説明することを困難にするが，より厳密に Set1 を特徴づけることができる．したがって，状況に応じて適当な Set2 を選ぶことが必要であるといえる．

5. おわりに

本稿では，部分グラフを制約とするグラフ構造データからの知識獲得の一手法として，専門家の持つ領域知識の精緻化を提案した．また，精緻化を実現するにあたり，B-GBI 法を用いた部分グラフの同型問題の近似解法を提案した．予備実験の結果，提案手法を用いて獲得したパターンが知識の精緻化に有用であることを確認した．

今後，本稿において得られたパターンが示す知識が実際に有用であるかどうかを，未知データに対する予測精度や，専門家の評価に基づき検証することが必要である．また，パターンの形式や探索制御に関する制約などを導入することで，より有用なパターンの獲得を考えたい．

参考文献

[Fortin 96] S. Fortin. Technical Report 96-20, University of Alberta, Edomonton, Alberta, Canada. The graph isomorphism problem, 1996.

[Quinlan 86] J. R. Quinlan: Induction of decision trees, *Machine Learning*, Vol.1, pp.81-106, 1986.

[Matsuda 02] T. Matsuda, H. Motoda, T. Yoshida, and T. Washio: Knowledge Discovery from Structured Data by Beam-wise Graph-Based Induction, in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence, Springler Verlag, LNAI2417*, pp. 255-264, 2002.

[松田 01] 松田，元田，鷲尾：一般グラフ構造データに対する Graph-Based Induction とその応用，人工知能学会誌，Vol.16，No.4，pp.363-374，2001.

[松田 02] 松田，元田，吉田，鷲尾：Graph-Based Induction による分類学習のための構造データからの属性構築，2002 年度人工知能学会全国大会（第 16 回）論文集，セッション 1A4-03 , 2002.

[吉田 97] 吉田，元田：逐次ペア拡張に基づく昨日推論，人工知能学会誌，Vol.12，No.1，pp.58-97，1997.

[山口 02] 山口高平：慢性肝炎データセットのクレンジングとマイニングの試み，情報洪水時代におけるアクティブマイニングの実現，平成 13 年度科学研究費補助金 特定領域 (B) 研究成果報告書，pp.205-221，2002.