1A2-4

グラフ構造データからの特徴的なパターン抽出における制約に基づく探索制御

Constrained Search Strategy in Extracting Discriminative Patterns from Graph Structured Data

高林 健登 Kiyoto Takabayashi　Phu Chien Nguyen　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 Chunkingless Graph-Based Induction (Cl-GBI) can extract typical patterns from graph structured data by the operation called chunkingless pairwise expansion which generates pseudo-nodes from selected pairs of nodes in the data. Cl-GBI enables to extract overlapping subgraphs, while its time and space complexities could be extremely high. Because of the complexities, it happens that Cl-GBI cannot extract large patterns within limited time and computational resources. In that case, extracted patterns could not be good enough to describe characteristics of data and not be so much of interest for domain experts. To mine more discriminative patterns which could not be extracted by the current Cl-GBI, we propose a search strategy using domain knowledge or interests of domain experts as constraint on patterns. Furthermore, we experimentally evaluate the proposed method using the hepatitis dataset, and show that it can extract more discriminative patterns.

1. はじめに

大量に蓄積された電子化データから興味深い有用な知識を獲得するデータマイニングにおいて，近年，複雑な構造を有するデータを扱うためにグラフ構造データを対象としたグラフマイニングが活発に研究されている [Yoshida 95, Matsuda 02, Kuramochi 04]．その一手法である Graph Based Induction（GBI）法 [Yoshida 95] は，ノードペアを逐次拡張（チャンク）することにより，グラフ中に頻繁に現れる典型的なパターンを高速に発見することができる．また，GBI 法のチャンキング時のあいまい性およびチャンクすることによる探索空間の不完全性などの問題を軽減した Beam-wise GBI（B-GBI）法 [Matsuda 02] も提案されている．しかしながら，GBI 法および B-GBI 法は部分的に重複するパターンを同時に抽出できなかったため，筆者の所属する研究グループではその問題を解消した Chunkingless GBI（Cl-GBI）法 [Nguyen 05] を提案した．Cl-GBI 法では，ノードペアをチャンクせずに一つの塊として捉えること（擬似チャンキング）で重複パターンの抽出を可能としたが，空間計算量，及び時間計算量が急激に増加する傾向にあった．そのため，限られた計算時間，及び計算資源の下ではパターンが大きく成長せず，抽出されたパターンが探索対象データの特徴を十分に現すことができなかった．

そこで本稿では，領域知識や興味深い部分パターンを制約として用いて Cl-GBI 法の探索を制御する手法を提案する．提案手法では領域知識，もしくはデータ解析者の興味に基づいた部分パターンを制約として与え，それらを含むパターンのみを抽出する，もしくは含まないパターンのみを抽出するように探索を制御することにより，限られた時間と計算資源の下において従来の手法では抽出できなかったパターンを抽出する．なお，従来の手法では特徴的なパターンをデータベースのグラフ中に頻出するパターンと定義していたが，本稿では，あるパターンを決定木などに用いたときにクラス分類性能が高いパターンと定義し，クラス分類性能の指標として information gain [Quinlan 86] を用いる．また，以下ではノードペアのことを単にペアと呼ぶ．

連絡先: 高林 健登　〒 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 高次推論方式研究分野
電子メール:kiyoto_ra@ar.sanken.osaka-u.ac.jp

2. Chunkingless Graph-Based Induction（Cl-GBI）法

2.1 Cl-GBI 法の概要

図 1 は，入力グラフ中のノード 1,2，及び 3 からなる典型的なパターンが擬似チャンキングにより抽出される過程を示している．具体的には，まず入力グラフ中のノード 1 と 3 からなるペアが擬似チャンクされ，擬似ノード 10 として登録される．その後，擬似ノード 10 とノード 2 からなるペアが擬似チャンクされることで擬似ノード 11，すなわち前述の典型的なパターンが抽出される．

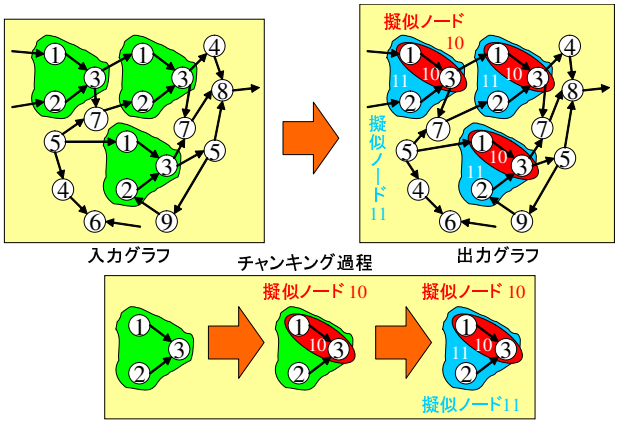

図 1: 擬似チャンキングの基本概念

また，Cl-GBI 法ではある一定の数（ビーム幅）だけペアを選択し，それぞれのペアについて擬似チャンクする．そうすることで典型的なパターンとなり得るペアが探索範囲からもれてしまう可能性を軽減することができる．

以下に Cl-GBI 法のアルゴリズムを示す．Cl-GBI 法は，ビーム幅 b，擬似チャンキングの繰り返しの最大数 N，及びペアが満たす最低支持度 θ をパラメータとして持ち，これらによって探索空間を制御する．言い換えるなら，各繰り返しでは最低支持度が θ 以上のペアの中から b 個が選択され擬似チャンクさ

れる．この各繰り返しをレベルと呼ぶ．理論的には，θ を 0 にし，b と N を十分に大きく設定することで，Cl-GBI 法は可能な全ての部分グラフを抽出できる [Nguyen 05]．

Cl-GBI 法のアルゴリズム

Input. グラフデータベース D，ビーム幅 b，最大レベル N，頻度の閾値 θ(%)

Output. 典型的なパターンの集合 S（初期値は空集合）

Step 1. D 中のグラフから隣接する 2 つのノードから成る全てのペアを抽出する．レベル 2 以降については，2 つのノードのうち少なくとも一方は新しく登録された擬似ノードからなるペアの全てを抽出する．

Step 2. 抽出されたペアの頻度を数える．ここで，θ よりも低い頻度のペアは削除する．

Step 3. **Step 1.** で抽出されたペアの中から頻度の高い順に b 個のペアを選び，それぞれを抽出パターンとして S に加える．この時，ペアを構成するノードが擬似ノードであれば元のパターンに復元してから S に加える．擬似チャンクすべきペアがない場合，もしくは，レベルが N の場合はここで終了する．

Step 4. **Step 3.** で選ばれたペアにそれぞれ新しいラベルを割り当てる．ただし，グラフは書き換えない．そして，**Step 1.** に戻る．

3. 制約に基づいた逐次ペア拡張

3.1 制約に基づく探索制御の導入

従来の Cl-GBI 法では対象領域における既知の知見は全く考慮せずに大規模なデータベースの中から盲目的に頻出パターンを抽出していたため，対象領域における興味深いパターンとは関係の無いパターンを大量に抽出しており，時間計算量や空間計算量が膨大になっていた．データ中に頻繁に現れるパターンはデータの特徴を現す典型的なものであるが，クラス分類性能の高いパターンや対象領域の専門家にとって興味深いパターンは必ずしも頻度が高いとは言えず，それらのパターンの頻度がそれほど高くない場合，大量に抽出される頻出パターンは限られた時間と計算資源の下では目的とするパターンの抽出を妨げることになる．実際に，千葉大学医学部付属病院からご提供頂いた肝炎データ [山口 02] を Cl-GBI 法を用いて解析した結果，対象領域の専門家から見て興味深いパターンが必ずしも探索されないという指摘を受けている [茂木 05]．

そこで本研究では，領域知識や興味を制約として与えて領域知識や興味とは無関係なパターンを排除して探索を行うことで，従来の手法では抽出困難であった対象領域におけるより興味深いパターンや特徴的なパターンを抽出する手法を提案する．具体的には，本稿では以下の二通りの制約を導入した．

制約 1. 領域知識から特徴的な事象であると考えられるパターンや興味のある事象を表すパターンを “**含みたいパターン（INpattern）**” として与える．この場合，図 2（a）のようにこれを含むものを抽出するよう探索を制御する．

制約 2. 領域知識から特徴的な事象には無関係であることが明確なパターンや興味のない事象のパターンを “**含みたくないパターン（EXpattern）**”として与える．この場合，図 2（b）のようにこれを含むものを抽出しないよう探索を制御する．

なお，INpattern が指定された場合，INpattern に現れるノードラベル，もしくはリンクラベルを含むペアだけを擬似チャンキングの候補として列挙する．これにより，INpattern を全く含む可能性のないパターンを効率よく探索から排除できる．ただし，対象データにおいてノードラベル，もしくはリンクラベルのいずれかの出現頻度が高い場合は，パターンを十分に制限できなくなるため，当該ラベルに関係なく，他方のラベルを含むペアだけを擬似チャンキングの候補として列挙する．

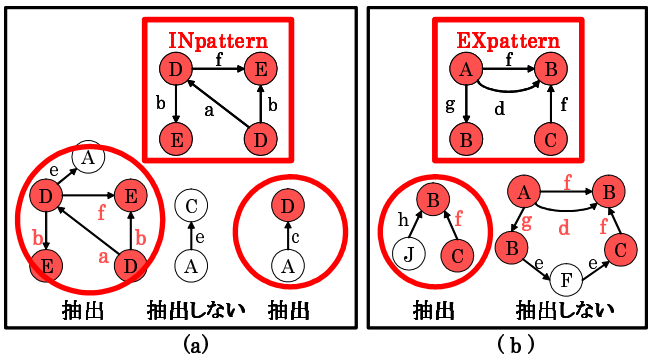

図 2: 制約の与え方の例

3.2 制約付き探索のアルゴリズム

制約に基づいて探索を制御するにあたって，抽出しようとしているパターン（抽出パターン）が制約パターンを含んでいるかを確認するとき，あるパターンが他のパターンに含まれるかを判定する部分グラフ同型問題を解くことは不可避である．しかしながら，部分グラフ同型問題は一般に NP 完全問題であり，可能な限り回避すべきであるため，この問題を解く回数をいかに減らすかが提案手法においては重要なポイントとなる．そこで本手法では，抽出パターンのノード情報，及びリンク情報から抽出パターンが制約パターンを含む可能性があるかをまず確認し，抽出パターンが制約パターンを含み得ないと判断できる場合には部分グラフ同型問題は解かず，そうでない場合にのみ部分グラフ同型問題を解く．ノード情報，及びリンク情報を用いた制約パターンの包含確認方法は次のとおりである．まず，抽出パターン P_i と制約として与えるパターン群 T の内の 1 つである制約パターン $T_j \in T$ について，式（1）で定義する T_{num} を比較し，$T_{num}(P_i, T_j)$ が $T_{num}(T_j, T_j)$ 未満ならば制約パターンを含む可能性がないので包含確認を終了する．

$$T_{num}(x, y) = \sum_{L_k \in L(y)} f(x, L_k) \tag{1}$$

ただし，$L(y)$ はパターン y に現れるラベルの集合であり，$f(x, L_k)$ はパターン x 中に現れるラベル L_k の数

そうでなければ，次に T_j と P_i について，式（2）で定義する P_{info} を比較し，$P_{info}(P_i, T_j)$ が $P_{info}(T_j, T_j)$ 未満ならば制約パターンを含む可能性がないので包含確認を終了する．

$$P_{info}(x, y) = \sum_{L_k \in L(y)} p(x, y, L_k) \tag{2}$$

ただし，

$$p(x, y, L_k) = \begin{cases} 1 & f(x, L_k) \geq f(y, L_k) \text{ の場合} \\ 0 & \text{そうでない場合} \end{cases}$$

もし，$P_{info}(P_i, T_j) \geq P_{info}(T_j, T_j)$ ならば，P_i が制約パターン T_j を含んでいる可能性があるため，部分グラフ同型問題を解いて T_j を含んでいるか否かを確認する．

図 3，4 に，制約に基づく探索制御を導入した Cl-GBI 法のアルゴリズムの内，探索制御を行う部分であり，従来の Cl-GBI 法において **Step 1.** にあたる部分のアルゴリズムを示す．INpattern を制約として用いる場合のアルゴリズムを図 3 に，Expattern を制約として用いる場合のアルゴリズムを図 4 に

記述する．どちらも入力はグラフデータベース D，制約パターンの集合 T，レベル Lv，及び抽出されたペアのリスト L で，出力は新たに抽出されたペアを加えた L である．また，アルゴリズム中の $PD(P_i, T_j)$ は部分グラフ同型問題を解く手続きであり，P_i が T_j を含んでいれば true を返し，含んでいなければ false を返す．具体的には，まず P_i 中の T_j に現れないノードとリンクのセットを削除し，次にこの前処理された部分グラフと T_j を Cl-GBI 法で同時に入力し，T_j 自身が抽出されるまで探索する．最終的に，T_j 自身以外に T_j と同じパターンが抽出されたか否かで，P_i に T_j が含まれるかどうかを判定する．

```
ExtPair(D, T, L, Lv)
input: database D, a set of constraint pattern T, Level Lv,
       a list of extracted pairs L (initial entry is empty set)
output: L with newly extracted pairs added
 begin
  if Lv = 1     then
        Enumerate pairs in D, which consist of nodes or links
        appearing in T, and store them in E_pin
  else     Enumerate pairs, which consist of one or both
        pseudo-nodes in L, and store them in E_pin
  for all P_i ∈ E_pin
  begin
   if P_i is not marked     then
      if T_num(P_i, T) ≥ T_num(T_j, T_j)     then
         if P_info(P_i, T_j) ≥ P_info(T_j, T_j)     then
            if PD(P_i, T_j) = true   then   mark P_i
            L := L ∪ {P_i}
         else     L := L ∪ {P_i}
      else     L := L ∪ {P_i}
   else     L := L ∪ {P_i}
  end
  return L
 end
```

図 3: 制約パターンが INpattern のときのパターン抽出アルゴリズム

なお，図 3 のアルゴリズムについては，制約パターンを完全に包含するパターンだけではなく，制約パターンを部分的にしか含まないパターン（近傍パターン）をも抽出されることに注意されたい．これは，INpattern を指定する際，用いる制約パターンが必ずしも目的とするパターンの一部である必要は無く，その一部を含んでいればよいことを意味する．

4. 探索制御を用いた評価実験

4.1 実験設定

本実験では，千葉大学医学部付属病院からご提供頂いた慢性肝炎データセットを用いた．慢性肝炎データセットに関しては，文献 [Geamsakul 05] と同様にインターフェロン投与の効果があった患者のクラスを R（Response），効果のなかった患者のクラスを N（Non-response）として，24 個の検査項目を属性として用いた．本実験で用いた慢性肝炎データセットのグラフサイズを表 1 にまとめる．なお，慢性肝炎データセットのグラフ構造データへの変換の詳細は文献 [Geamsakul 05] を参照されたい．

4.2 実験方法

INpattern と EXpattern の 2 種類が利用可能な Cl-GBI 法を計算機（CPU: Pentium 4 3.2 GHz，Memory: 4GB，OS: Fedora Core release 3）上に C++を用いて実装し，慢性肝炎データセットに適用した．ただし，同時に利用可能な制約は INpattern，EXpattern のいずれかである．本実験では INpattern の制約のみを利用するが，慢性肝炎データセットにおいてノードラベルは検査項目に共通であり頻出するため，INpattern のリンクラベルのみを抽出パターン列挙の際の制約として用いた．また，提案手法は前述の通り制約を完全に満足しない近傍パターンも探索できることから，制約として機能することがある程度期待できる部分パターンを制約パターンとして利用した．具体的には，文献 [沼尾 05] を基にして図 5 の 3 つのパターン群をそれぞれ INpattern として用い，Cl-GBI 法のパラメータの内，ビーム幅 b を 10 に，擬似チャンキングの最大繰り返し回数 N を 10 に，頻度の閾値 θ を 0 に固定し，計算時間，及び抽出されたパターンの information gain の最大値を観測した．また，同条件で制約を指定せずに Cl-GBI 法を実行した際に抽出されたパターンの内，最も information gain の高かったパターン（MDpattern）を制約パターン（領域知識）として使用した場合の計算時間と，抽出されたパターンの information gain の最大値も同様に観測した．用いたパターンを図 6 に示す．ここで，図 5，6 の制約として与えるパターンは，ある時点での検査結果を表したもので，ラベル d を持つノードは時間軸上のある時点を表すダミーノードである．た

```
ExtPair(D, L, T, Lv)
input: database D, a set of constraint pattern T, Level Lv,
       a list of extracted pairs L (initial entry is empty set)
output: L with newly extracted pairs added
 begin
  if Lv = 1     then     Enumerate all the pairs in D
                         and store them in E_pex
  else     Enumerate pairs, which consist of one or both
        pseudo-nodes in L, and store them in E_pex
  for all P_i ∈ E_pex
  begin
   if T_num(P_i, T_j) ≥ T_num(T_j, T_j)     then
      if P_info(P_i, T_j) ≥ P_info(T_j, T_j)     then
         if PD(P_i, T_j) = true   then   discard P_i
         else     L := L ∪ {P_i}
      else     L := L ∪ {P_i}
   else     L := L ∪ {P_i}
  end
  return L
 end
```

図 4: 制約パターンが EXpattern のときのパターン抽出アルゴリズム

表 1: 慢性肝炎データセットのグラフのサイズ

クラス	R	N
グラフ数	38	56
平均ノード数	104	112
最多ノード数	145	145
最少ノード数	24	20
ノード数の合計	3,944	6,296
ノードラベル数	12	
平均リンク数	108	117
最多リンク数	154	154
最少リンク数	23	19
リンク数の合計	4,090	6,577
リンクラベル数	30	

とえば，図 5 の No.1 中の最も左のパターンは，ある時点の検査結果のうち GPT の値が H で PLT の値が L であったことを意味する．

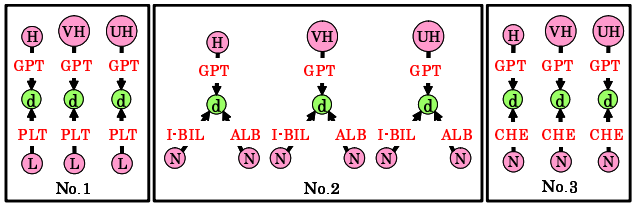

図 5: 実験で用いた INpattern

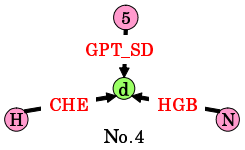

図 6: 従来の手法で抽出された MDpattern

4.3 実験結果と考察

実験結果を表 2 に示す．表 2 は制約を用いなかった場合（original），及び 4 種類の制約を用いた場合（No.1〜No.4）のそれぞれに対して計算時間，抽出パターン中の information gain の最大値（max information gain）を示したものである．表 2 において max information gain の欄に着目すると，No.1 と No.3 を制約として与えた手法は従来の手法よりも information gain の値が低いパターンを抽出しているが，No.2 と No.4 を制約として与えた手法では従来の手法よりも information gain の値が高いパターンを抽出している．また，No.3 以外の制約を与えた手法で抽出された MDpattern には図 7（a），(b)，(d) に示すように，図 5，6 の制約として与えたパターンの内の 1 つが含まれており，No.3 を制約として与えた手法に関しても図 7（c）のように与えた制約パターンの内の 1 つの部分パターンを含むものが MDpattern となっている．

表 2: 実験結果

	time	max information gain
original	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

以上のことから，与える制約が適切であれば，提案手法によって従来手法よりも特徴的なパターンを抽出できると考えられる．また，No.4 のような抽出された MDpattern を繰り返し制約パターンとして探索を行うことによって，さらに特徴的なパターンを抽出できることが予想できる．このことから，パターンの探索を繰り返しながら属性を構築しつつ決定木を構築する DT-ClGBI 法 [Nguyen 06] のようなアプリケーションにおいて，提案手法が属性生成手法として有効に機能することが期待できる．さらに，制約を与えた手法は従来の手法の 1/2〜1/6 の計算時間でほぼ同程度かそれ以上に特徴的なパターンを抽出しており，提案手法は与える制約が適当なものならば特徴的なパターンを効率的に探索できると考えられる．

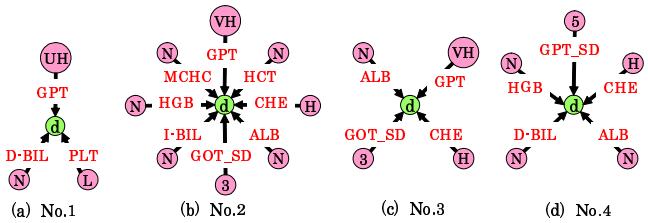

図 7: 実験で抽出された MDpattern

5. おわりに

本稿では，領域知識や興味に基づいた制約により探索を制御するグラフマイニング手法を提案した．また，評価実験により，提案手法は与える制約が対象領域において適当なものであれば，盲目的に頻出パターンを抽出する従来手法よりもクラス分類性能の高いパターンを抽出できることを示した．提案手法は従来手法によって抽出されたクラス分類性能の高いパターンを制約パターンとして用いることでよりクラス分類性能の高いパターンを抽出できる．

今後の課題としては，制約の与え方を INpattern と EXpattern のどちらかではなく，両者を同時に指定することでより柔軟な制約の指定方法の実現，及びパターンの探索を繰り返しながら決定木を構築する DT-ClGBI 法における提案手法の属性生成手法としての有効性の検証が挙げられる．

参考文献

[Yoshida 95] K. Yoshida and H. Motoda. *CLIP: Concept Learning from Inference Patterns*. Artificial Intelligence, Vol. 75, No. 1, pp. 63-92, (1995).

[Matsuda 02] T. Matsuda, H. Motoda, T. Yoshida, and T. Washio. *Mining Patterns from Structured Data by Beam-wise Graph-Based Induction*. Proc. of DS 2002, pp. 422-429, (2002).

[Kuramochi 04] M. Kuramochi and G. Karypis. *An Efficient Algorithm for Discovering Frequent Subgraphs, IEEE Trans*. Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1038-1051, (2004).

[Nguyen 05] P.C. Nguyen, K. Ohara, H. Motoda, and T. Washio. *Cl-GBI: A Novel Approach for Extracting Typical Patterns from Graph-Structured Data*. Proc. of PAKDD 2005, pp. 639–649, (2005).

[山口 02] 山口高平，畑澤寛光，佐藤芳紀. **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**. 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, 情報洪水時代におけるアクティブマイニングの実現, pp. 205–221, (2002).

[茂木 05] 茂木明，Phu Chien Nguyen, 大原剛三，元田浩，鷲尾隆. *DT-ClGBI* **法による肝炎データからの知識発見**. 人工知能学会研究会資料, SIG-KBS-A405, pp. 19–25, (2005).

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

[Geamsakul 05] W. Geamsakul, T. Yoshida, K. Ohara, H. Motoda, H. Yokoi, and K. Takabayashi. *Constructing a Decision Tree for Graph-Structured Data and its Applications*. Fundamenta Informaticae Vol.66, No.1-2, pp. 131–160, (2005).

[沼尾 05] 沼尾正行，市瀬龍太郎，佐藤慶宜，本山真也. **インターフェロンの効果予測**. 平成 13 年度〜平成 16 年度科学研究費補助金　特定領域研究研究成果報告書，情報洪水時代におけるアクティブマイニングの実現，pp. 273–279, (2005).

[Nguyen 06] P.C. Nguyen, K. Ohara, A. Mogi, H. Motoda, and T. Washio. *Constructing Decision Trees for Graph-Structured Data by Chunkingless Graph-Based Induction*. Proc. of PAKDD 2006, pp. 390–399, (2006).