2F3-01

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探索の効率化

Improving the Efficiency of Extracting Typical 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 discriminative patterns from graph structured data by the operation called chunkingless pairwise expansion which generates pseudo-nodes from selected pairs of nodes in the data. The extracting patterns as pseudo-nodes are useful for constructing a classifier such as a decision tree. But the space and time complexities of Cl-GBI could be extremely high because the number of pseudo-nodes explosively increases as the search progresses. To improve the search efficiency of Cl-GBI, we introduce a heuristic criterion to select which pairs to be pseudo-chunked. For that purpose, we first analyze the search tendency of generated patterns by conducting preliminary experiments. Then, we design a heuristic function to evaluate pairs of nodes which is able to shift the weights of two kinds of criteria according to the progress of the search. Furthermore, we show the results of experimental evaluations using a promoter dataset and a hepatitis dataset, and discuss the usefulness of the proposed method.

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特徴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 [Chien 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，その反面，擬似ノードが探索の進行とともに指数的に増加するため空間計算量・時間計算量が爆発的に増加する傾向にあった．

そこで本稿では，特徴的なパターンを効率的に抽出するために有効なペアを擬似チャンキングの対象として選定する新たな選定基準（評価関数）を提案し，Cl-GBI 法の探索を効率化する．なお，本稿で特徴的なパターンと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を意味し，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．

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 および 3 からなる特徴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

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 元田研究室
電子メール:kiyoto_ra@ar.sanken.osaka-u.ac.jp

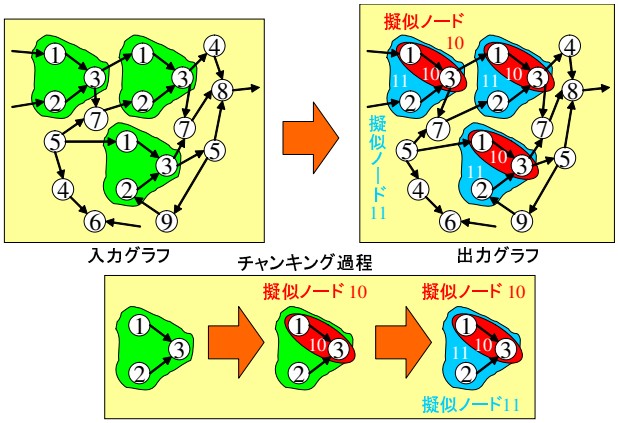

図 1: 擬似チャンキングの基本概念

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，すなわち前述の特徴的なパターンが抽出される．

また，Cl-GBI 法では，擬似チャンクするペアを唯一ではなくある一定の数（ビーム幅）だけ選択し，それぞれのペアについて擬似チャンクする．そうすることで特徴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，およびペアが満たす最低支持度 θ をパラメータとしてもち，これらにより探索空間が制御される．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，最低支持度を 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 [Chien 05]．

Cl-GBI 法のアルゴリズム

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

output. 特徴的なパターンの集合 S（初期値は空集合）

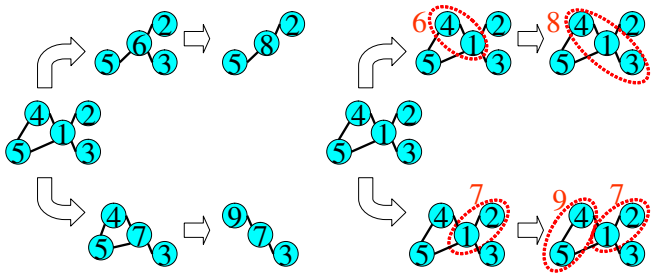

図 2: 擬似チャンキングによる計算量の増加

Step 1. D のグラフ中の隣接する二つのノードから成る全てのペアを抽出し，それらの頻度を数える．レベル 2 以降については，二つのノードのうち少なくとも片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し，頻度を数える．ここで，θ よりも低い頻度のペアは擬似チャンクすべきペアとして数えることに意味を持たないので削除する．

Step 2. 頻度の高い順に b 個のペアを **Step 1.** で抽出されたペアの中から選び，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．この際，擬似チャンクすべきペアがなければ終了する．また，レベルが N の場合もここで終了する．

Step 3. **Step 2.** で選ばれたペアを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．

2.2 Cl-GBI 法における探索に関する問題点

Cl-GBI 法は，選択したペアを 1 つの新ノードに置き換えて（チャンクして）元のグラフを書き換える GBI 法・B-GBI 法と異なり，擬似ノードを生成するのみで元のグラフを書き換えない．これにより，Cl-GBI 法では部分的に重複するパターンを同時に抽出できるが，その反面，擬似チャンキングによって考慮すべきノード数が増加することにともない，GBI 法や B-GBI 法ではチャンクする度に減少していたチャンクするペアの組み合わせが指数的に増加し，その結果，空間計算量と時間計算量が爆発的に増加している．

例として図 2 を考える．図 2（a）は B-GBI 法の b=2，N=3 のチャンキング過程を，図 2（b）は Cl-GBI 法の b=2，N=3 の擬似チャンキング過程を示している．B-GBI 法は，Cl-GBI 法と同様にビーム幅で指定された b 個のペアをチャンクするが，グラフを書き換える必要があるため状態を b 個に分割し，並列にチャンキングを進めるグラフマイニング手法である．図 2（b）における点線で囲まれたノード群は擬似ノードを表している．図 2（a），(b）を比較してまず気づくことは，B-GBI 法ではノード数が減少しているのに対して Cl-GBI 法ではノード数が減少せず，逆に擬似ノードの分だけ増加している点である．このことからも探索空間が広く，B-GBI 法よりも Cl-GBI 法のほうが計算量が多いことがわかる．表 1 に B-GBI 法の各レベルにおけるチャンキング候補，表 2 に Cl-GBI 法の各レベルにおける擬似チャンキング候補であるペアを列挙する．

表 1 では状態を増やした直後に一旦増加したチャンキング候補のペアが，次の時点では減少しているのに対し，表 2 では

表 1: B-GBI 法のレベルごとのチャンキング候補のペア

レベル	チャンキング
0	1-2, 1-3, 1-4, 1-5, 4-5
1	2-6, 3-6, 5-6, 3-7, 4-5, 4-7, 5-7
2	2-8, 5-8, 3-7, 7-9

表 2: Cl-GBI 法のレベルごとの擬似チャンキング候補のペア

レベル	擬似チャンキング
0	1-2, 1-3, 1-4, 1-5, 4-5
1	1-2, 1-3, 1-4, 1-5, 4-5, 1-2-3, 1-2-4, 1-2-5, 1-3-4, 1-4-5
2	1-2, 1-3, 1-4, 1-5, 4-5, 1-2-3, 1-2-4, 1-2-5, 1-3-4, 1-4-5, 1-2-3-4, 1-2-4-5, 1-3-4-5

レベルが進むにつれて擬似チャンキングの候補ペアが単調に，かつ大幅に増加していることがわかる．

3. ヒューリスティックスによる逐次ペア拡張の効率化

3.1 Cl-GBI 法の探索効率の傾向分析

クラス分類性能の指標として用いる information gain は，チャンキングによるパターンの成長過程に対して非単調に変化するため，単純に information gain を擬似チャンキングの際のペアの選定基準に採用してもクラス分類性能の高いパターンへと成長するペアを選定できるとは限らない．したがって，Cl-GBI 法の探索を効率化する上で，まず Cl-GBI 法の探索がどのような傾向にあるのかを認識することは大きな意味を持つ．そこで Cl-GBI 法の探索傾向を検証するために予備実験を行い，その結果から Cl-GBI 法の探索効率を向上させる要素を選定するというアプローチを取る．

予備実験においては，Cl-GBI 法のビーム幅 b を 10 に，頻度の閾値 θ を 0 に固定し，N を 1〜10 まで変化させた場合に対して，擬似チャンクするペアの選定基準として頻度を用いた場合（以下 $p1$ と呼ぶ）の探索傾向を調べた．具体的には，計算時間と探索の進行状況，および抽出されるパターンのクラス分類性能の関係を調べた．さらにペアの選定基準に information gain を直接用いて同様にパラメータを変化させた場合（以下 $p2$ と呼ぶ）の結果も検討した．なお，入力データには UCI Repository の Machine Learning Database [Blake 98] から取得した promoter データセットを用いた．計算時間と最大 infomation gain の関係を図 3 に示す．

$p1$ と $p2$ を比較した結果，$p1$ は $p2$ よりも特徴的なパターンを抽出できるが，その反面，膨大な計算時間を要し，逆に $p2$ では計算時間は大幅に少なくて済むが特徴的なパターンが探索範囲に含まれない傾向にあることがわかった．また，パターンの種類に関しては $p1$，$p2$ ともにレベルが進むにつれて計算量を増加させる原因となる特徴的なパターンを形成するペアに関与しないパターンが増える傾向にあった．

実験結果の計算時間と抽出されたパターンの種類の関係に注目すると，$p1$ は特徴的なパターンを抽出するとともに，そうでないものも多く抽出しているのに対して $p2$ は抽出しているパターンの種類が極端に少なかった．抽出されるパターンの

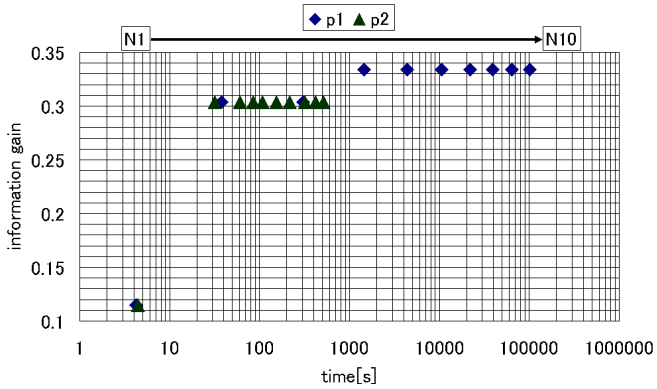

図 3: 計算時間と最大 information gain の関係

種類が少なければ探索過程で考慮すべきペアの組み合わせも減少し，計算時間も短くて済む．なぜ $p2$ の抽出するパターンの種類が少なくなるのかの原因は，information gain を擬似チャンクするペアの選定基準とした場合，途中経過において得られた擬似ノードを用いて生成するペアがすでに抽出済みのパターンと重複する割合が，頻度を擬似チャンクするペアの選定基準とした場合よりも非常に高いためと考えられる．つまり，information gain をペアの選定基準とした場合，類似するペアが選ばれる傾向にあり，本来の探索空間の特定の領域を重点的に探索するのに対し，頻度をペアの選定基準とした場合は，多様なペアが選択される傾向にあり，本来の探索空間の広範にわたって探索を進めていると考えられる．

3.2 ヒューリスティック関数の設計

あるパターンが探索の初期レベルで抽出されないということは，その部分パターンの評価値がその時点で低かったことを意味し，探索が相当進まない限りそのパターンを抽出することは困難である．したがって，初期レベルでは対象とするペア数がそれほど多くなく計算量的にも負荷が大きくないことから，可能な限り網羅的に探索を進めることが重要となる．一方，探索が進むにつれて擬似チャンキングの性質上，探索空間が飛躍的に拡大し，計算量も指数的に増加する傾向にあるため，探索後半に関しては計算量を抑えるために網羅的ではなく特徴的なパターンを限定的に探索する必要がある．

したがって，探索初期ではペアの選定基準として頻度を，探索後半では information gain を用いることで探索を効率化できると考えられる．

以上の分析結果に基づき，擬似チャンクするペアの優先度を制御する関数を設計する．前述の議論にしたがい，探索の初期レベルにおけるペアの選定基準として頻度を採用し，レベルが進行するにつれて information gain をペアの選定基準として採用するために，まず以下のような関数を考える．

$$H(e) = \alpha \times F(e) + \beta \times I(e) \tag{1}$$

ただし，　$\alpha \geq 0$，　$\beta \geq 0$，　$\alpha + \beta = 1$ であり，$F(e)$ はペア（部分パターン）e の頻度を表し，$I(e)$ は e の information gain を表している．α と β はそれぞれ $F(e)$ と $I(e)$ に対する重みであり，これらの値は探索の進行とともに変化する必要があるため，ともにレベル L に対する関数でなければならない．そこで，予備実験の結果に見られる傾向を考慮してそれぞれを以下のように定義した．

$$\alpha = f(L) = \frac{\exp(-L^2)}{\exp(-L^2) - \log(\frac{N-L}{N})^2} \tag{2}$$

$$\beta = g(L) = -\frac{\log(\frac{N-L}{N})^2}{\exp(-L^2) - \log(\frac{N-L}{N})^2} \tag{3}$$

上記の式において，L はレベル，N は繰り返し回数（Cl-GBI 法のパラメータ）を意味し，レベルは 0〜$N-1$ までの値をとる．直観的には，$\alpha(L)$ は 1 から始まり，レベルが進行するにつれて 0 に漸近していく関数で，$\beta(L)$ は 0 から始まり，レベルが進行するにつれて 1 に漸近していく関数である．

4. ヒューリスティック関数を用いた評価実験

4.1 実験設定

本実験では，UCI Repository の Machine Learning Database から取得した promoter データセットと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て実験を行った．慢性肝炎データセットに関しては，文献 [Yoshida 04] 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．各データセットのグラフサイズをそれぞれ表 3，4 にまとめる．なお，各データセットのグラフ構造データへの変換の詳細はそれぞれ文献 [Geamsakul 03] および [Yoshida 04] を参照されたい．

表 3: promoter データセットのグラフのサイズ

クラス	promoter	non-promoter
グラフ数	53	53
グラフ一枚中のノード数	57	57
ノード数の合計	3021	3021
ノードラベル数	4	
グラフ一枚中のリンク数	515	515
リンク数の合計	27295	27295
リンクラベル数	10	

4.2 実験方法

本実験では，擬似チャンクするペアの選定基準に，グラフ中のパターンの出現頻度（以下 $p1$ と呼ぶ）と，information gain（以下 $p2$ と呼ぶ），および 3.2 節で設計した評価関数の値（以下 $p3$ と呼ぶ）の 3 種類が利用可能な Cl-GBI 法を計算機（CPU: Pentium III 930 MHz，Memory: 512 MB）上に C++を用いて実装し，promoter および慢性肝炎データセットに適用した．具体的には，Cl-GBI 法のパラメータのうち，ビーム幅 b を 10 に，頻度の閾値 θ を 0 に固定し，繰り返し回数 N を 1〜10 の範囲で変化させ，各選定基準に関して，計算時間，抽出されたパターンの種類，抽出されたパターンの information gain の最大値を観測した．計算時間と information gain の最大値，および抽出されたパターンの関係を調べ，探索の効率と特徴的なパターンの抽出能力を評価する．ただし，$p1$ を用いた慢性肝炎データに対する実験においては，$N = 8$ 以降でメモリがオーバーフローしたため，N を 1〜7 までの範囲で変化させた場合の結果のみを評価対象とした．

4.3 実験結果と考察

promoter データセットの結果を図 4 に，慢性肝炎データセットの結果を図 5 に示す．図 4 および図 5 はそれぞれの計算時間と抽出パターン中の information gain の最大値の関係を示したものである．

図 4 において $p3$ と $p1$ を比較すると，$N = 4$〜10 の範囲で $p1$ の 1/100 の計算時間で同じ特徴的なパターンを抽出している．また，図 4 において $p3$ と $p2$ を比較すると，$p3$ は $N = 4$〜10 の範囲で $p2$ よりもわずかに多くの計算時間をかけているものの，$p2$ よりも特徴的なパターンを抽出している．

表 4: 慢性肝炎データセットのグラフのサイズ

クラス	N	R
グラフ数	56	38
平均ノード数	112	104
最多ノード数	145	145
最少ノード数	20	24
ノード数の合計	6296	3944
ノードラベル数	12	
平均リンク数	117	108
最多リンク数	154	154
最少リンク数	19	23
リンク数の合計	6577	4090
リンクラベル数	30	

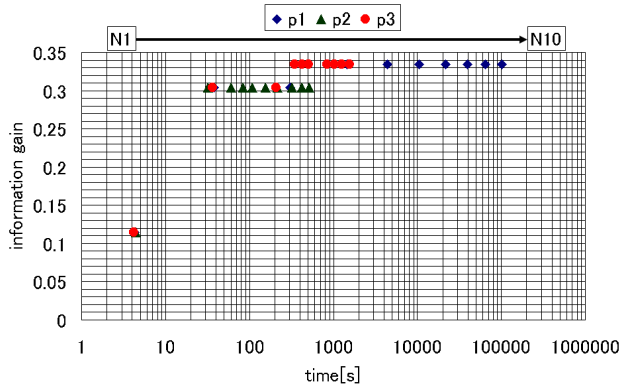

図 4: promoter における計算時間と最大 information gain の関係

この結果は初期レベルにおいて頻度を重視することでパターンを網羅的に抽出し，後半において空間計算量の削減のために information gain を重視したことが最大限に成功したことを示していると考えられる．以上の結果から，promoter データセットに関して $p3$ は非常に効率的な探索をしていると考えられる．

次に，図 5 において $p3$ を $p1$ と比較すると，$p3$ は $p1$ よりも短時間で特徴的なパターンを抽出している．また，図 5 において $p3$ を $p2$ と比較すると，$p3$ は $p2$ ほど計算時間が短くないが，$N = 6, 7$ において最も特徴的なパターンを抽出している．これは初期レベルにおいて網羅的にパターンを抽出していたことによって得られた結果ではないかと考えられる．

また，各最大繰り返し回数のときの計算時間と抽出されたパターンの種類の関係に注目すると，promoter データセットに関しては $p2$，および $p3$ はほぼ同じ計算時間で同程度のパターンを抽出していたが，$p1$ だけは非常に多くの計算時間をかけて大量にパターンを抽出していた．これに対して慢性肝炎データセットに関しては $p1$，$p3$，$p2$ の順に計算時間が多くなり，それに応じて抽出パターン数も多くなったが，promoter データセットと比較すると，計算時間，抽出パターン数のいずれにおいても $p1$，および $p2$，$p3$ 間で promoter データセットのように大きな差が開くような傾向は見られなかった．したがって，用いるデータセットによって $p1$，および $p2$，$p3$ それぞれの抽出パターン数の傾向に違いがあることになる．これはデータの特性によるものであると考えられる．慢性肝炎データセットのノードラベル，およびリンクラベルは promoter データセットのそれよりも圧倒的に多く，このことが 3.1 節で述べた抽出パターンの重複を少なくし，抽出パターン数に大きな違いが生じなかった原因ではないかと考えられる．その場合，$p1$ が網羅的にパターンを抽出する傾向にあることから，$p1$ は特徴的なパターンを抽出しきれずに探索を終えていると解釈でき，また，$p3$ が $p2$ よりも特徴的なパターンを抽出するのに時間がかかっているのも同じ理由であると解釈できる．

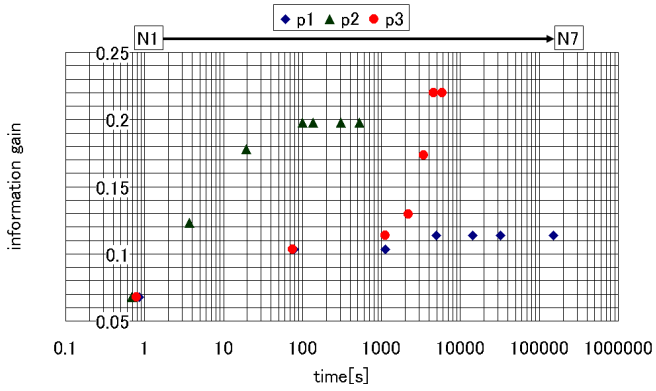

図 5: 肝炎データにおける計算時間と最大 information gain の関係

このように，本稿で提案した評価関数をペアの選定基準として実装した Cl-GBI 法はノードラベルやリンクラベルの比較的少ないデータに対しては非常に効率的な探索を行うことが可能だが，ノードラベルやリンクラベルの比較的多いデータに対しては計算時間を要求する．しかしながら，頻度をペアの選定基準とする従来の Cl-GBI 法と比較すると格段に計算時間を短縮することに成功しており，Cl-GBI 法の探索効率の改善という点においては良好な結果を得ることができたと言える．

5. おわりに

本稿では，Cl-GBI 法における探索を効率化するために，擬似チャンクするペアの優先度を探索過程に応じて制御するペアの評価関数を提案し，評価実験によりその有効性を検証した．提案した評価関数は，探索初期では頻度を，探索が進むにしたがって information gain を評価基準として主として用いるというヒューリスティックスに基づいており，頻度のみを用いていた従来手法と比較して，同程度のクラス分類性能をもつパターンを非常に高速に抽出できる．一方，評価実験を通して，提案した評価関数を用いてもノードラベルやリンクラベルの多いデータに関しては，それらが少ないデータよりも多くの計算時間を要することが明らかとなった．

今後の課題としては，データの構造・特性が既知である人工データを用いて提案した評価関数を擬似チャンキングの際の選定基準として採用した Cl-GBI 法のより詳細な性能解析を進めるとともに，データの構造・特性による影響がより少ない評価関数，もしくは探索手法を検討する必要がある．

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).

[Chien 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, (to appear).

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

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

[Blake 98] C.L. Blake, E. Keogh, and C.J. Merz. UCI Repository of Machine Learning Database. http://www.ics.uci.edu/˜mlearn/MLRepository.html, (1998).

[Geamsakul 03] W. Geamsakul, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. *Performance Evaluation of Decision Tree Graph-Based Induction*. Proc. of DS 2003, pp. 128-140, (2003).

[Yoshida 04] T. Yoshida, W. Geamsakul, A. Mogi, K. Ohara, H. Motoda, T. Washio, H. Yokoi, and K. Takabayashi. *Preliminary Analysis of Interferon Therapy by Graph-Based Induction*. Proc. of AM 2004, pp. 31-40, (2004).