3F2-03

# グラフマイニングとILPシステムの比較考察

Comparison of Graph Mining Method with Inductive Logic Programming Systems

猪口 明博[*1]
Akihiro Inokuchi

[*1]日本アイ・ビー・エム株式会社 東京基礎研究所
Tokyo Research Laboratory, IBM Japan

Some approaches which can efficiently discover characteristic patterns from a set of labeled graphs or trees have been proposed. Authors proposed a method which achieves an efficient search of all frequent subgraph patterns from labeled graphs, and extended it to discover more meaningful patterns. In this paper, we consider the relation between patterns found by our methods and descriptions of graphs represented by predicates of first order logic.

## 1. はじめに

近年，ラベル付きグラフや順序木，木の集合などから頻出するパターンを効率良く抽出する手法が提案されている．筆者らは，ラベル付きグラフの集合から頻出するパターンを効率良く列挙するAGMアルゴリズム [Inokuchi 03a] を提案し，さらにより有用なパターンを取り出すために，AGMアルゴリズムを拡張した．本稿では，それらの手法によって取り出されるパターンと一階述語で記述されるグラフの知識表現の関連について考察する．

## 2. グラフマイニング

### 2.1 多頻度グラフパターンの抽出

グラフとはラベルをもつ頂点と辺からなる構造で，多頻度グラフ抽出問題とは，グラフの集合が与えられたときにこれらのグラフの多くに部分グラフとして含まれる部分構造を見つける問題である．図1は，多頻度グラフ抽出の例を示したものである．入力として4つのグラフと最小支持度として100%が与えられたとき，出力されるグラフパターンの1つとして，頂点数が3のグラフが出力される．アイテム集合の場合と同じく，頂点数（あるいは辺数）が少ないグラフパターンから探索していき，徐々に頂点数（辺数）が多いパターンを探索するアルゴリズムが提案されている．代表的な手法としてはAGM [Inokuchi 03a], FSG, gSpan, Gaston, FREQT, TreeMinerなどがある．詳細は [浅井, Nijssen] を参照されたい．

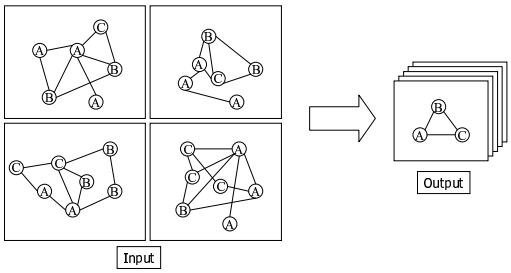


図 1: 頻出パターン抽出問題

### 2.2 グラフパターンペアの抽出

2.1節で述べた頻出パターンとして取り出される部分グラフが必ずしもより良い知識表現であるとは限らない．例えば，図2の正例の破線部分が，ある薬理活性に大きく寄与する場合を考える[*1]．頻出パターン抽出手法が活性の有無を分類するために適用された場合，図2のパターン1に示される$P_1$は取り出されないかもしれない．それは，パターンの定義は部分グラフであり，パターン$P_1$が図2の負例の破線部分を含む活性が無い化合物にも含まれているからである．従って，図2の正例の破線部分を意味するパターンを取り出すための知識表現が必要となる．筆者らは，部分グラフパターン$P_1$を含むが$P_2$を含まないグラフは活性がある ($P_1 \wedge \neg P_2 \Rightarrow$ *positive*) というルールを効率良く取り出す手法を提案した．提案手法は，単に後処理でパターンの組合せを考えるのではなく，探索の途中で部分グラフパターンのペアの評価値の上限を見積もりながら探索するという手法である．

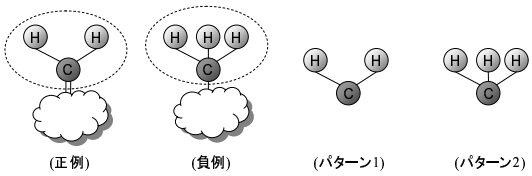


図 2: グラフの例

### 2.3 汎化パターンの抽出

アイテム集合を対象とした問題では，アイテム間の階層概念を考慮した相関ルールを求める手法が既に提案されている．筆者はグラフの頂点ラベル，辺ラベルに階層概念が存在するときにグラフデータ集合に頻出する部分グラフを取り出す手法を提案した [猪口 04]．考慮するデータがアイテム集合の場合，集合中のアイテムに重複がない場合としているが，グラフの場合，同じラベルをもつ頂点や辺がグラフの中に複数存在するので，アイテム集合の場合に比べて，多頻度グラフの数が膨大になりうる．また図3のように分子構造をグラフとして考え，炭素 (C) と窒素 (N) の上位概念をAとし，最小支持度100%で全ての多頻度グラフを取り出すと，頂点数6のパターンは図3の12個のパターンが得られる．ここで，左上以外のパターンは過度に一般化されていると見なせ，必要なパターンは左上の1つである．後処理として不必要なパターンを除く手法は非効率であるので，[猪口 04] の手法は，探索の途中で，過度に一般化された部分グラフパターンを枝狩りし，効率化を図っている．

連絡先: 神奈川県大和市下鶴間 1623-14, Tel:046-215-4343, Fax:46-273-7428, inokuchi@jp.ibm.com

*1 図2の負例の破線部分は活性が有る化合物だけでなく，無い化合物にも含まれているものとする

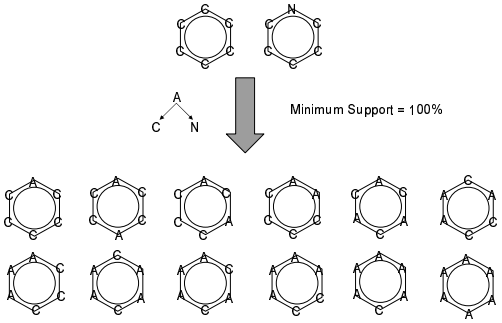


図 3: 汎化パターンと過度の汎化

## 3. 一階述語論理によるグラフの表現

[Inokuchi 03b] によって取り出されるルールは，図 4 のパターン $P_1$ を含むが，$P_2$ を含まないデータは正例 (あるいは，負例) であるというルールである．これを，述語 $node(a, b_1, c)$ と $link(a, b_1, b_2, c)$ で表すと以下のようになる．

$$node(1, 1, O) \wedge node(1, 2, C) \wedge link(1, 1, 2, d)$$
$$\wedge \neg node(1, 3, Cl) \wedge \neg link(1, 2, 3, s)$$

ここで，$a$ はグラフ ID，$b_i$ は頂点 ID，$c$ は頂点ラベル，あるいは辺ラベルである．即ち，[Inokuchi 03b] で得られるルールは，否定 (negation) の項 (term) を含む述語文に相当する．

一方，[猪口 04] によって取り出されるパターンは，図 4 のパターン $P_4$ のようなパターンであり，述語で表すと以下のようになる．

$$node(1, 1, O) \wedge node(1, 2, C) \wedge link(1, 1, 2, d)$$
$$\wedge node(1, 3, x) \wedge link(1, 2, 3, s)$$

ここで，$x$ は変数である．即ち，[猪口 04] で得られるルールは，変数を含む述語の項を含む述語文に相当する．さらに $x$ がハロゲン原子だとすると，$is\text{-}a(x, halogen)$ という述語を使うことで，変数を階層化して表すことができる．

グラフマイニングによって限量子 (quantifier) を含むパターンを取り出すことはできないが，グラフマイニングによって取り出されるパターンと一階述語で記述されるグラフの知識表現はほぼ等価であることが分かる．

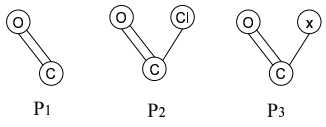


図 4: 述語とグラフマイニングのパターンの関係

## 4. 実験

本節では，実データを用いて，グラフマイニングと ILP システムの比較実験を行う．用いたデータは，変異原性の有無が分かっている 188 個 (正例:125，負例:63) の化合物で，原子と結合をそれぞれグラフの頂点と辺として扱い，原子の種類と結合の種類をそれぞれ頂点ラベルと辺ラベルとして扱った．手法 [Inokuchi 03b, 猪口 04] を AcGM(Apriori-based connected Graph Mining) アルゴリズムに統合し，評価実験を行った．AcGM アルゴリズムはラベル付きのデータ集合から頻出する連結グラフパターンを効率良く取り出す手法である．実験は [Hoche 03] と同じ 10-fold cross validation を行い，正答率と標準偏差を表 1 に示す．ラベルの階層は図 5 を用いた．

表 1: グラフマイニングと ILP システムの比較 [Hoche 03]

| methods | Acc. | Stand. Dev. |
|---|---|---|
| ACQ | 87.0 | n/a |
| $C^2$RIB | 88.0 | ±3.4 |
| FOIL | 82.0 | ±3.0 |
| Fors | 89.0 | ±6.0 |
| G-Net | 92.0 | ±8.0 |
| Progol | 88.0 | ±2.0 |
| STILL | 90.0 | ±5.0 |
| BACQ(250) | 92.0 | ±3.8 |
| AGM | 89.0 | ±2.0 |

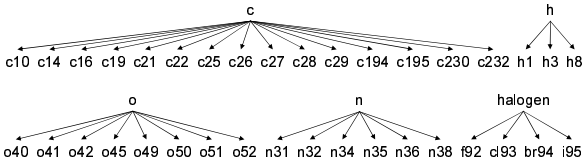


図 5: 実験に用いたラベルの階層

## 5. むすび

本稿では，筆者らが提案したグラフマイニングの手法によって取り出されるグラフパターンと一階述語で記述される知識表現について考察を行った．今後，筆者らの手法と ILP の手法のアルゴリズムの関連や大規模データへの適用の可能性について考察する予定である．

## 参考文献

[浅井] 浅井，有村．半構造データマイニングにおけるパターン発見技法，電子情報通信学会論文誌 D-I，Vol.J87-D-I, No.2，pp. 79–96，2004.

[Hoche 03] Hoche, S. Horvath, T. & Wrobel, S. Effective Rule Induction from Molecular Structures Represented by Labeled Graph *Proc. of Int'l Workshop on MGTS*, 2003.

[古川] 古川，尾崎，植野，帰納論理プログラミング，共立出版

[Inokuchi 03a] Inokuchi, A., Washio, T and Motoda, H. Complete Mining of Frequent Patterns from Graphs: Mining Graph Data. *Machine Learning*, Vol. 50 Issue. 3, pp. 321–354, 2003.

[Inokuchi 03b] Inokuchi, A. & Kashima, H. Mining Significant Pairs of Patterns from Graph Structures with Class Labels. *Proc. of Int'l Conf. on Data Mining*, pp. 83–90, 2003.

[猪口 04] 猪口．ラベルの概念階層を利用したグラフマイニング第 64 回知識ベースシステム研究会．pp. 259–264, 2004.

[Nijssen] http://hms.liacs.nl/index.html