平成 24 年度　　修士論文概要

関係的知識発見におけるパターンの表現に関する研究

犬塚研究室　西尾 典晃
学籍番号　No. 23417594

## 1　はじめに

データマイニングの目的は，データ中に潜む有用な知識を発見することである。データ中に頻繁に出現するパターンを枚挙することを特に頻出パターン枚挙とよぶ。関係的知識発見 (MRDM) は，複数の関係表に跨るパターンを発見する。MRDM は帰納論理プログラミング (ILP) の枠組みで行われる。これは論理的な記述によって豊かな表現力を持つ。

本研究では論理性に基づいた有意味なパターンの表現について 2 つの観点からアプローチした。一つは，パターンを分析する手法である形式概念分析 [4] の MRDM への適用について議論し，簡潔な論理パターンを得る提案である [5]。形式概念分析では数学的性質から特定のパターン (形式概念) を抽出するが，[5] ではその形式概念を決定する論理的包含関係に基づく極小パターンを求めた。この提案は，意味内容に言及したパターンの選択を行っているといえる。

Warmr [1] は頻出な論理パターンを枚挙するが，データによっては現実的な時間での枚挙が困難である。Mapix [2] は意味のある論理式のまとまりをアイテムとし，ボトムアップに探索する。これはトップダウンにパターンを探索する Warmr に比べ格段に高速な枚挙を可能にしている。ところで，これらの手法が対象としているデータは構造物のような一つ一つが区別された事例からなる。このようなデータ以外に，事例が互いに関係しあう開いたネットワーク型のデータも存在する。もう一つの提案は，この開いた構造を持つデータを対象とした，アイテムの生成方法及びボトムアップにパターンを探索する手法である。

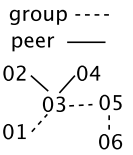
## 2　ILP における頻出パターン枚挙

ILP の枠組みでは関係 rel のタプル $\langle t_1, \ldots t_n \rangle$ を，論理式 $\mathsf{rel}(t_1, \ldots, t_n)$ として表現する。また探索空間を制御するため述語の引数に入力 $(+)$/出力 $(-)$ の情報を与える。図 1 のデータベース $\mathbf{r}_{\mathrm{net}}$ はあるネットワークを表しており，頂点を表す関係 member(以下 m) と頂点間のリンクを表す関係 peer$(+,-)$ と group$(+,-)$ からなる。このとき，m を目標述語，その基礎原子式を事例，および事例の含む基礎項を目標項とよぶ。例えば m(05) は事例であり，05 は目標項である。

パターンは後件が目標述語，前件がそれ以外の述語の連言で構成される次のような節である。

$$S_1 = \mathsf{m}(A) \leftarrow \mathsf{group}(A,B) \wedge \mathsf{peer}(B,C) \wedge \mathsf{group}(B,D)$$
$$S_2 = \mathsf{m}(P) \leftarrow \mathsf{peer}(P,Q) \wedge \mathsf{group}(P,R) \wedge \mathsf{group}(R,S)$$
$$S_3 = \mathsf{m}(X) \leftarrow \mathsf{peer}(X,V) \wedge \mathsf{group}(X,W) \wedge \mathsf{group}(W,Y) \wedge \mathsf{group}(X,Z)$$

ここで $S_3\theta \subseteq S_2$ を満たす置換 $\theta = \{X/P, V/Q, W/R, Y/S, Z/R\}$ が存在する。このとき $S_3$ は $S_2$ を包摂する ($S_3 \succeq S_2$) という。同様に $S_2 \succeq S_3$ が成り立つ。




| member |
|---|
| 01 |
| ... |
| 06 |

| group | |
|---|---|
| 01 | 03 |
| 03 | 05 |
| 05 | 06 |

| peer | |
|---|---|
| 02 | 03 |
| 03 | 04 |

図 1: 開いた構造を持つデータベース $\mathbf{r}_{\mathrm{net}}$; 左図は $\mathbf{r}_{\mathrm{net}}$ が現わす構造を模式的に表したグラフ

このとき $S_2$ と $S_3$ は同値であるという。またパターン $p$ の頻度 (支持度) とは，$p$ を満たす事例の割合をいう。ILP データマイニングにおける頻出パターン枚挙とは，与えられた最低支持度以上の支持度を持つ同値でない頻出パターンをすべて枚挙することである。

## 3　開いた構造対象の頻出パターン枚挙

**開いた構造**　従来の手法が対象としていたデータの事例は互いに独立しているが，事例間がつながっている特徴を示すデータも存在する。本論文では前者を閉じた構造，後者を**開いた構造**とよび，区別する。閉じた構造の例として，Bongard データセットでの事実

$$\mathsf{bongard}(2) \leftarrow \mathsf{circle}(2,\mathsf{o2}) \wedge \mathsf{triangle}(2,\mathsf{o3}) \wedge \mathsf{in}(2,\mathsf{o2},\mathsf{o3})$$

を挙げる。これは絵中に丸と三角があり，丸の中に三角があることを意味する。Bongard の事例はどれもこのように，絵に描かれている図形に関する情報を持つ。ところが，複数の絵が一つ事例に含まれていたり，事例が含む図形がまた事例になっている，ということはない。これは事例が一つの構造物であり，事実はその部分構造を表現していることに起因する。つまり，直感的には工業製品や分子構造等と同様に，その構造物に関して閉じている。以上より，事例は互いに独立しているといえる。対して開いた構造はこれとは対照的な特徴を示す。例えば$\mathbf{r}_{\mathrm{net}}$は，目標項が複数の関係に現われており，事例 m(03) についての事実

$$\mathsf{m}(03) \leftarrow \mathsf{peer}(03,02) \wedge \mathsf{peer}(03,04) \wedge \mathsf{group}(03,05)$$

について，これ自身が含む定数項はさらにそれ自身が事例になり得る。以下の節では，開いた構造に対して頻出パターンを枚挙する手法について述べる。

### 3.1　Hanabi

表 1 の Hanabi アルゴリズムは開いた構造を持つデータベースを入力として，事例の近傍からアイテムを生成し，その頻出な組合せを枚挙する。以下にアイテムとなる花火節とその組合せ手続きの概要を示す。

**花火節**　まず，花火節を生成するための近傍について述べる。ある事例 $e$ の持つ近傍とは，以下のいずれかを満たすリテラルの集合 $N$ である。

⋆ $e$ の項を含む

⋆ 入力引数と出力引数両方に近傍基礎項を含む

ここで近傍基礎項とは $e$ の項を入力引数に含むような基礎リテラルが出力引数に含んでいる定数項の集合で

ある。つまり，事例 m(03) に関する近傍は

$N_{03} = \mathsf{peer}(03, 04) \wedge \mathsf{group}(03, 05)$

という事実である。そして $var(e \leftarrow N)$ のように変数化し，**花火節**を得る。事例 m(03) に関する花火節は

$s_{03} = \mathsf{m}(A) \leftarrow \mathsf{peer}(A, B) \wedge \mathsf{group}(A, C)$

のようになる。なお，花火節において前件に含まれるリテラルの出力項の集合の中で，入力項に現れていない項を連結項とよぶ。

**重ね合わせ**　閉じた構造とは異なり，同一の花火節の組合せでもパターンとしては複数あり得る，さらに同じ花火節同士を組合せて別のパターンを作ることができる。$s_{03}$ は連結項として $B$ と $C$ を持つので，$s_{03}$ 同士でも組合せ方は以下の 3 通りある。

(1) $\mathsf{peer}(A, B) \wedge \mathsf{group}(A, C) \wedge \mathsf{peer}(B, D) \wedge \mathsf{group}(B, E)$
(2) $\mathsf{peer}(A, B) \wedge \mathsf{group}(A, C) \wedge \mathsf{peer}(C, D) \wedge \mathsf{group}(C, E)$
(3) $\mathsf{peer}(A, B) \wedge \mathsf{group}(A, C) \wedge \mathsf{peer}(B, D) \wedge \mathsf{group}(B, E)$
$\wedge \mathsf{peer}(C, D) \wedge \mathsf{group}(C, E)$

この組合せ情報を保存するため，**パターン木**を導入する。パターン木は，ノードに花火節の ID を持ち，花火節の組合せ方に従って子を持つ木である。組合せ情報を保存することで，Apriori での下方閉包性 [3] を開いた構造に応用し効率的に頻出パターンを枚挙できる。表 1 の SpShell 手続きはパターン木の集合を入力として，すべての可能な重ね合わせを出力する。

**実験**　パターン数の比較を行った。テストデータは頂点数 12，リンク数 17 のネットワークである。表 2 は最低支持度 1/12 において，重ね合わせを行わない場合 (基本的組合せ) と，行う場合 (重ね合わせ) にそれぞれ枚挙された候補パターン数を表している。三行目は頻出パターン数である。基本的組合せでは，4-花火パターン以降は組合せが厖大となりシステムが停止している。重ね合わせにより候補数を大幅に小さくできていることがわかる。

表 1: パターン木重合せ及び頻出パターン枚挙

SpShell($\mathcal{T}_k$):
**input** : パターン木の集合 $\mathcal{T}_k$;
**output** : 重ね合わせ $\mathcal{T}_{k+1}$;
1. $\mathcal{T}_{k+1} := \emptyset$;
2. **for each** $T \in \mathcal{T}_k$ **do**
3. $\mathcal{S} := \emptyset$;
4. $Sub^T :=$ $T$ から根を除いて得られる木の集合;
5. **for each** $Sub_i^T \in Sub^T$ **do**
6. $s := \emptyset$;
7. **for each** $T_j \in \mathcal{T}_k$ **do**
8. **if** $Sub_i^T$ と最深の葉を除いた $T_j$ が同型
9. **then** $s := s \cup \{\langle Sub_i^T, T_j \rangle\}$;
10. $\mathcal{S} := \mathcal{S} \cup \{s\}$;
11. $\mathcal{T}_{k+1} := \mathcal{T}_{k+1} \cup \{\textsc{SpGen}(T, \mathcal{S})\}$;
12. **return** $\mathcal{T}_{k+1}$;

SpGen($T$, $\mathcal{S}$):
**input** : パターン木 $T$; 置換ノード $\mathcal{S}$;
**output** : $T$ の重ね合わせ;
1. $\mathcal{N} := \emptyset$;
2. **for each** $S_i \in \mathcal{S}$ **do**
3. **select** $s \in S_i$;
4. $\mathcal{N} := \mathcal{N} \cup s$; $S_i := S_i \setminus s$;
5. SpGen($T$, $\mathcal{S}$);
6. **for each** $\langle subtree, t \rangle \in \mathcal{N}$ **do**
7. **substitute** $subtree$ in $T$ with $t$;
8. **return** $T$;

Hanabi ($\mathbf{r}$, $\sup_{\min}$):
**input** : データベース $\mathbf{r}$; 最低支持度 $\sup_{\min}$;
**output** : 頻出花火パターン $Freq$;
1. $\mathcal{U} := \emptyset$; $k := 1$;
2. $\mathcal{U} :=$ $\mathbf{r}$ の事例に関する花火節;
3. $\mathcal{F}_1 := \{S \in \mathcal{U} \mid S$ は頻出 $\}$;
4. **while** $\mathcal{F}_k \neq \emptyset$ **do**
5. $\mathcal{C}_{k+1} :=$ SpShell($\mathcal{F}_k$のパターン木);
6. $\mathcal{F}_{k+1} := \{CS \in \mathcal{C}_{k+1} \mid CS$ は頻出 $\}$;
7. $Freq := Freq \cup \mathcal{F}_{k+1}$; $k := k + 1$;
8. **return** $Freq$;

表 2: $\sup_{\min} = 1/12$ でのパターン数の比較

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 基本的組合せ | 4 | 76 | 6,512 | - | - | - |
| 重ね合わせ | 4 | 56 | 843 | 47,881 | 35,760 | 0 |
| 頻出 | 4 | 26 | 317 | 7,735 | 10,848 | 0 |

## 4　おわりに

本論文は MRDM におけるパターンの表現に関して提案した。[5] はパターンの論理表現に着目し，その冗長性を除く手法について提案した。また，開いた構造を持つデータの特徴についても明らかにし，そのような構造を持つデータに対し，頻出パターンを枚挙する手法を与えた。今後の課題として，開いた構造を持つデータに関する具体的検討や，Hanabi について様々なデータを用いた適用実験が挙げられる。

### 参考文献

[1] Dehaspe, L. and Toivonen, H. Discovery of frequent datalog patterns. *Data Min. Kowl. Disov.*, Vol.3, No.1, pp.7-36, 1999.

[2] Nakano, Y. and Inuzuka, N. Multi-relational pattern mining based-on combination of properties with preserving their structure in examples. *ILP'2010*, *LNCS*, Vol.6489, pp.181-189, 2011.

[3] Agrawal, R. et al. Mining Association Rules between Sets of Items in Large Databases. *SIGMOD Conference'1993. ACM Press*, pp.207-216, 1993.

[4] Wille, R. Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. *LNCS*, Vol.3626, pp.1-33, 2005.

### 発表論文

[5] Nishio, N., Mutoh, A. and Inuzuka, N. On Computing Minimal Generators in Multi-Relational Data Mining with respect to $\theta$-Subsumption. *ILP2012*.