2B-5

開いた構造を持つ事例を対象とした関係的知識発見

西尾 典晃 †　　犬塚 信博 †
† 名古屋工業大学大学院 工学研究科 情報工学専攻

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。これは述語論理形式でパターンを表現する手法で，豊かな表現力を持つが計算コストが大きい。本論文では既存の ILP 手法があまり扱っていなかった構造のデータにおいて，頻出パターン枚挙を可能にする手法について提案する。

2　ILP における頻出パターン枚挙

初期に提案された Warmr [1] は Apriori [3] と同様にレベルワイズに頻出パターンを枚挙する。Mapix [2] はアイテムをボトムアップに組合せて頻出パターンを枚挙することにより，Warmr を高速化した。しかしながら，既存手法は親子関係などの閉じた構造を前提としており，ネットワークのような事例間の切れ目が明らかでない構造 (開いた構造) はあまり考慮されておらず，適用が難しい。

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 $rel(t_1, \ldots, t_n)$ 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 に示すデータベース$\mathbf{r}_{sample}$はあるネットワークを表しており，頂点を表す関係 person(以下 p) と頂点間のリンクを表す関係 peer$(+,-)$ と conf$(+,-)$ からなる。このとき，person 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 p(05) は事例であり，05 は目標項である。

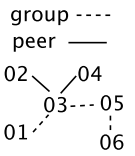
パターンは後件が目標述語，前件がそれ以外の述語の連言で構成される次のような節である。

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

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。このとき S_2 と S_3 は同値であるという。次にパターンの頻度について導入する。パターン S の支持度とは，S を満たす事例の割合を示す。いま，与えられたデータベース $\mathbf{r}$，目標項の集合 $\mathbf{t}$ における S の支持度を

$sup(S, \mathbf{r}, \mathbf{t}) = |\{e \in \mathbf{t} \mid S(e) \text{ succeeds w.r.t. } \mathbf{r}\}| \ / \ |\mathbf{t}|$

と定義する。頻出パターン枚挙とは，与えられた最低支持度 $\sup_{\min}$ 以上の支持度を持つ同値でないパターンをすべて枚挙することである。



person	conf		peer	
01	01	03	03	04
...	03	05	03	02
06	05	06		

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

3　開いた構造対象の頻出パターン枚挙

提案手法はネットワークを表すデータベースを入力として，事例の近傍から基本アイテムを生成し，その頻出な組合せをパターンの重ね合わせにより枚挙する。以下に花火アイテムとその組合せ手続きの概要を示す。

花火アイテム　まず，花火アイテム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，以下のいずれかを見たすリテラルの集合 N である。

⋆ e の基礎項を含む
⋆ 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 の基礎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 p(03) に関する近傍は

$N_{03} = \mathsf{p}(03) \leftarrow \mathsf{peer}(03,02), \mathsf{peer}(03,04), \mathsf{conf}(03,05).$

という事実である。いくつかの事例の近傍を変数化したリテラル集合において包摂関係における同値類の中で極小なリテラル集合を花火アイテムという。事例 p(03) に関する花火アイテムは以下のようになる。

$S_{03} = \mathsf{p}(A) \leftarrow \mathsf{peer}(A,B), \mathsf{conf}(A,C).$

グラフの観点から，花火アイテムは注目頂点とその頂点から一ステップで辿れる頂点の集合により与えられるネットワークの誘導部分グラフに相当する。なお，花火アイテム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。

パターンの重ね合わせ　花火アイテムは通常のアイ

A Multi-Relational Mining Method for Open Structure Examples
†Noriaki NISHIO　†Nobuhiro INUZUKA
†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

テムセットの場合と異なり，同一のアイテム集合でもアイテム間の繋がりは複数存在するため，それぞれは異なったパターンとなる。ここでアイテム間の連結情報を保存しておくため，パターン木を導入する。パターン木とは頂点ラベルに花火アイテムの ID を持つ順序木である。各々の頂点は対応する花火アイテムの連結項の数だけ，その連結項が出現したリテラルの順序に従って子ノードを持つことができる。深さ d のパターン木を持つ花火パターンを $(d+1)$-花火パターンと呼ぶ。

ここで k-頻出花火パターンから効率的に $(k+1)$-頻出花火パターンを枚挙することを考える。最も単純には k-頻出花火パターンの任意の連結項に花火アイテムを連結し，$(k+1)$-候補パターンとして生成する。しかしこれは頻出でないと分かっているパターンも候補として生成してしまうため，重ね合わせにより効率的に候補パターンを生成する。

パターン S の P との重ね合わせ S' とは，S のパターン木 T_S からルートを除いてできる木の集合 Sub^{T_S} の元 $Sub_i^{T_S}$ と，P のパターン木 T_P から深さが極大の葉ノードを除いた T'_P が同型であるとき，T_S の $Sub_i^{T_S}$ と P を置換した木が表す 2 パターンである。例えば先のパターン S_1, S_2 の重ね合わせ S'_1 は，次のようなパターンである。

$S'_1 = \mathsf{p}(A) \leftarrow$
$\mathsf{conf}(A,B), \mathsf{peer}(B,C), \mathsf{conf}(B,D), \mathsf{conf}(D,E).$

表 1 SuperpositionShell 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。

4 実験

提案手法の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，重ね合わせを行う場合 (表 1 Hanabi) 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を示している。基本的組合せの場合，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。

5 まとめ

本論文では，開いた構造を対象とした ILP の枠組みでのパターン枚挙の手法を提案した。その際，効率的に候補パターンを生成するために重ね合わせを導入した。今後の課題として，より大きな規模のデータに対して適用の検討が必要である。

表 1: 花火パターン重合せ及び頻出パターン枚挙

SuperpositionShell($\mathcal{T}_k$):
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. $\quad \mathcal{S} := \emptyset$;
4. $\quad Sub^T := T$ からルートを除いて得られる木の集合;
5. $\quad$ **for each** $Sub_i^T \in Sub^T$ **do**
6. $\quad\quad s := \emptyset$;
7. $\quad\quad$ **for each** $T_j \in \mathcal{T}$ **do**
8. $\quad\quad\quad$ **if** Sub_i^T と最深の葉を除いた T_j が同型 **then**
9. $\quad\quad\quad\quad s := s \cup \{\langle Sub_i^T, T_j\rangle\}$;
10. $\quad\quad \mathcal{S} := \mathcal{S} \cup \{s\}$;
11. $\quad \mathcal{T}_{k+1} := \mathcal{T}_{k+1} \cup \{\mathrm{SpGen}(T, \mathcal{S})\}$;
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. $\quad$ **select** $s \in S_i$;
4. $\quad \mathcal{N} := \mathcal{N} \cup s$; $S_i := S_i \setminus s$;
5. SpGen(T, $\mathcal{S}$);
6. **for each** $\langle subtree, t\rangle \in \mathcal{N}$ **do**
7. $\quad$ **substitute** $subtree$ in T to t;
8. **return** T;

Hanabi ($\mathbf{r}$, $\mathbf{t}$, $\sup_{\min}$):
input : データベース $\mathbf{r}$; 目標事例 $\mathbf{t}$; 最低支持度 $\sup_{\min}$;
output : 頻出花火パターン $Freq$;
1. $\mathcal{U} := \emptyset$; $k := 1$;
2. **for each** $e \in \mathbf{t}$ **do** $\mathcal{U} := \mathcal{U} \cup e$ の花火アイテム;
3. $\mathcal{F}_1 := \{S \in \mathcal{U} \mid sup(S, \mathbf{r}, \mathbf{t}) \geq \sup_{\min}\}$;
4. **while** $\mathcal{F}_k \neq \emptyset$ **do**
5. $\quad \mathcal{C}_{k+1} :=$ SuperpositionShell($\mathcal{F}_k$のパターン木);
6. $\quad \mathcal{F}_{k+1} := \{CS \in \mathcal{C}_{k+1} \mid sup(CS, \mathbf{r}, \mathbf{t}) \geq \sup_{\min}\}$;
7. $\quad Freq := Freq \cup \mathcal{F}_{k+1}$; $k := k+1$;
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

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. and Srikant, R. Fast algorithms for mining association rules in large database. *VLDB'94*, *Morgan Kaufmann*, pp.487-499, 1994.