3K-01

OI包摂に基づくネットワークパターンマイニングアルゴリズムの効率的実装

森遼太† 武藤敦子† 森山甲一† 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

複数のテーブルからなるデータベースを対象とするデータマイニングに関係型データマイニング [1] がある．特に社会ネットワークを扱うアルゴリズムとしてHanabi[2] が提案されている．これは，Apriori 性に基づいて効率化が図られているが，膨大な数のパターンを生成する．そのため鷲見らは，包摂関係に基づき，支持度の計算を行わずに頻出を判定をする手法を提案した [3]．本論文ではこれを OI 包摂に対して拡張する．

2 Hanabi

Hanabi では，表 1 の形式を入力とする．単項の述語のリテラルを目標リテラルという．2 項の述語は対象間のネットワークを表現し，入出力の区別があるとする．

定義 1 (近傍) データベース D のある目標リテラル e の**近傍**は，次の (1)(2) で定まるリテラル集合である．
(1) 入力引数が e である D 内のリテラルは e の近傍である．このとき，その出力項を近傍基礎項という．
(2) 全ての引数が e の近傍基礎項である D 内のリテラルは近傍である．

ある e を頭部とし，その近傍を本体とした節を変数化したものを基本パターンという．例えば，表 1 の person1 から抽出される基本パターンは，$\mathrm{m}(x_1) \leftarrow \mathrm{f}(x_1, \boldsymbol{x_2})$, $\mathrm{f}(x_1, \boldsymbol{x_3})$, $\mathrm{f}(x_1, \boldsymbol{x_4})$, $\mathrm{f}(x_2, x_4)$, $\mathrm{f}(x_3, x_4)$. である．本体の出力変数を**連結変数**という．

定義 2 (支持度) ネットワークを表すデータベース D におけるパターン C の支持度は次の sup_c である．

$$sup_c = \frac{|\{t \in T | match(D, C, t)\}|}{|T|}$$

T は D の目標リテラルの集合である．支持度が最小サポートを超えるパターンを頻出パターンとする．

An efficient implementation for a Network Pattern Mining algorithm using OI-subsumption
Ryota Mori†, Atsuko Mutoh†, Koichi Moriyama†and Nobuhiro Inuzuka†
Dept. of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan†

表 1: 友人関係ネットワークに関するデータベース

m(X) (member)	f(X, Y) (friend)	
person1	person1	person2
person2	person1	person3
person3	person1	person4
person4	person2	person4
person5	person3	person4
person6	...	...

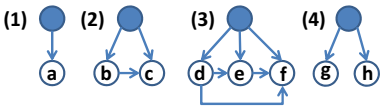

図 1: 基本パターン

定義 3 (OI 包摂) 花火節 C，データベース D，$t \in D$ について，$(1)t = \mathrm{head}(C)\theta$, $\mathrm{body}(C)\theta\rho \subseteq D$，$(2)\theta = \{A_1/b_1\}$, $\rho = \{A_2/b_2, \ldots, A_n/b_n\}$，$(3)i \neq j$ ならば $b_i \neq b_j$，を全て満たす代入 θ,ρ があるとき，C は t にマッチするといい，$match(D, C, t)$ と書く．

包摂はパターン同士も同様に定義される．θ 包摂では同じリテラルを何度でも代入することができ，$\mathrm{m}(x_1) \leftarrow \mathrm{f}(x_1, x_2)$, $\mathrm{f}(x_1, x_3)$. と $\mathrm{m}(x_1) \leftarrow \mathrm{f}(x_1, x_2)$. は同じパターンとなる．OI 包摂では異なる．

定義 4 (パターン木) データベース D の全基本パターンの集合を $\mathcal{U}$ とするとき，$\mathcal{U}$ のパターン木は次の通り．
$(1)P \in \mathcal{U}$ に対し，$\mathcal{U}$ の連結変数の個数だけ葉を持つ高さ 1 のラベル付き順序木 T は $\mathcal{U}$ のパターン木である．ただし T の根ノードは P 自身を，葉は P の連結変数をラベルとする．T は P を表す．
(2) $\mathcal{U}$ のパターン木 T と $P \in \mathcal{U}$ について，T のある葉 X のラベルを P に変更し，X に P を表すパターンを置換した木 T' もパターン木である．T' は T が表すパターン P_T と P に対し，$P_T \cup body(P)\theta$ を表す．ここで $\theta = \{x/head(P)\}$，x は X の元のラベルである．これを T に P を連結するという．

最後に，Hanabi のアルゴリズムを表 2 に示す．

表 2: ネットワークパターン枚挙アルゴリズム Hanabi

HANABI(D, $\sup_{\min}$):
input :データベース D; 最低支持度 $\sup_{\min}$;
output :頻出パターン $Freq$;
1. $\mathcal{U} := \emptyset$; $k := 1$; $D_t := D$ の目標リテラル;
2. **for each** $t \in D_t$ **do** $\mathcal{U} := \mathcal{U} \cup t$ の基本パターン;
3. $\mathcal{F}_1 := \{P \in \mathcal{U} \mid sup_P \geq \sup_{\min}\}$;
4. **while** $\mathcal{F}_k \neq \emptyset$ **do**
5. $\quad \mathcal{C}_{k+1} := \mathcal{F}_k$ にある基本パターンを連結;
6. $\quad \mathcal{F}_{k+1} := \{CS \in \mathcal{C}_{k+1} \mid sup_{cs} \geq \sup_{\min}\}$;
7. $\quad Freq := Freq \cup \mathcal{F}_{k+1}$; $k := k + 1$;
8. **return** $Freq$;

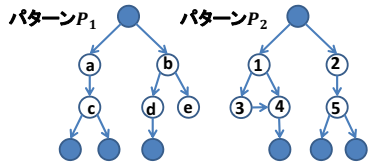

図 2: パターン同士の比較 (提案)

3 提案手法及び実験

鷲見らは，頻出でないパターンの拡大パターンは頻出でないことに基づいた提案を行った．また，基本パターン間の包摂関係を予め計算した上で再帰的にパターンを比較する方法を示した．一方 OI 包摂では図 1 の (4) のようなパターンが存在し，ノード g と h が対称なため，複数の代入パターンを考慮した方法が必要である．

まず，基本パターン同士について OI 包摂によるパターンマッチングを行い，包摂関係を確認すると共に可能な代入を全て得る．例えば図 1，(2) と (3) のマッチングでは，{d/b,e/c},{d/b,f/c},{e/b,f/c} という 3 通りの代入がある．再帰的に比較する際，代入のいずれかに対し，連結されている基本パターンが全て同じか拡大パターンになっていれば拡大パターンであると判定し候補から除外する．表 3 にアルゴリズムを示す．

例として，図 2 の P_2 が P_1 の拡大であることを調べる．まず，根の基本パターンはどちらも (4) であるため拡大の可能性がある．次に，その下に連結した基本パターンを比較する．ノード 1,2 への代入は，ρ_1 ={a/1,b/2} と ρ_2 ={a/2,b/1} の 2 通りがあり，その内 ρ_2 でパターン 2 側の基本パターンが拡大または同じとなる．さらにその下は，{c/5,d/4,e/3} でパターン 2 側の基本パターンが拡大または同じとなる．以上，最下まで達したため拡大と判定され，候補から除外される．

実験ではノード数 30，エッジ数 50 のランダムネットワークに，最小サポート 20%，OI 包摂によるマイニングを行った．実験の結果を表 4 に示す．マッチング回数は，パターンの支持度を計算した回数を表している．したがって，候補数 – マッチング回数が，提案手法によって候補から除外された数である．

表 3: パターン同士を再帰的に比較するアルゴリズム

Compare(P_1,P_2): %P_2 が P_1 の拡大パターンなら true
input : パターン木 P_1,P_2;
1. $B_1 \leftarrow P_1$ の根のラベルが表す基本パターン;
2. $B_2 \leftarrow P_2$ の根のラベルが表す基本パターン;
3. **if** (P_1 が葉ノード) **return** true; %予め計算済み
4. **if** (B_2 が B_1 の拡大か同じ) %計算済み
5. $\quad \rho_s \leftarrow B_1\theta\rho \subseteq B_2$ となる ρ の集合;
6. $\quad$ **foreach** $\rho \in \rho_s$ **do**
7. $\quad\quad \rho = \{X_1/Y_1, \cdots, X_n/Y_n\}$;
8. $\quad\quad$ **foreach** $(X_k/Y_k) \in \rho$ **do**
9. $\quad\quad\quad P_3 \leftarrow X_k$ から下の部分木;
10. $\quad\quad\quad P_4 \leftarrow Y_k$ から下の部分木;
11. $\quad\quad\quad$ **if** (Compare(P_3,P_4) == false) fail; 6 へ
12. $\quad\quad$ **return** true;
13. $\quad$ **return** false;
14. **return** false;

表 4: 最小サポート 20%:実験結果

木の高さ	候補数	頻出数	マッチング回数
1	9	3	9
2	1107	13	43
3	5517	38	89
4	7812	48	110
5	3582	51	107
6	2439	42	99
7	1998	36	82
8	1134	15	51
9	315	0	4

4 まとめ

候補パターン同士を，基本パターンを利用して再帰的に比較することでネットワークとのマッチング回数を減らす手法を OI 包摂に対して拡張した．提案手法の結果，OI 包摂を用いたパターンマイニングにおいてもマッチング回数を大幅に削減できることを確認した．しかしながらネットワークの規模をさらに大きくした際に，一部のパターンとネットワークとのマッチングで膨大な時間がかかってしまう問題が見つかっているため，ネットワークとパターンのマッチングについてのアルゴリズムを改善することが今後の課題である．

参考文献

[1] Luc Dehaspe and Hannu Toivonen. Discovery of frequent datalog patterns. *Data Min. Knowl. Discov.*, Vol. 3, No. 1, pp. 736, 1999.

[2] 西尾 典晃, 犬塚 信博. 開いた構造を持つ事例を対象とした関係的知識発見. 第 75 回情処全国大会, 2013.

[3] 鷲見 俊貴, 武藤 敦子, 犬塚 信博. ネットワークパターンマイニングにおける包摂関係を用いたアルゴリズムの改善. IPSJSIG Notes, ICS, 2015.