3U-3

基本パターンに基づく関係型パターンマイニング手法の効率性に関する考察

中野　裕介†　　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

データマイニングとは，データ中に潜む有用な知識を抽出する技法である．関係型データマイニング (MRDM) とは，複数の関係表にまたがるパターンを発見するアプローチである．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データ間の規則性を抽出する手法で，豊かな表現力を持つが計算コストが大きい．我々はこれまでに事例に現れる基本パターンの組み合わせに探索を限定する Mapix [2] を提案してきた．これは他の ILP 手法と比べて格段の計算速度でマイニングできることを示している．本稿では包摂関係に基づいたパターン生成の仕組みを提案し，これを導入することで効率的にマイニングできることを示す．

2 ILP データマイニング

ILP の枠組みでは関係 rel のタプル $\langle a_1, ..., a_n \rangle$ を，述語形式 $rel(a_1, ..., a_n)$ で表現しパターンをマイニングする．例えば図 1 の R_{fam} の関係表は $gf(X)$: X は祖父である，$p(X,Y)$: X は Y の親である，$m(X)/f(X)$: X は男性/女性である，と表現される．ここで祖父についてパターンを抽出したいとき，関係 gf を目標事例，また述語 $gf(X)$ を目標述語と呼ぶ．パターンは結論部が目標述語，前件部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．

$$gf(A) \leftarrow m(A) \wedge p(A,B) \wedge p(B,C) \wedge f(C).$$

ILP 手法の多くは探索空間を制御するために述語の引数に入力 (+)/出力 (−) の情報を与える．R_{fam} の述語には $p(+,-)$, $m(+)$, $f(+)$ とモードが与えられているとする．出力引数をもつ述語を経路述語，出力引数をもたない述語を判定述語と呼ぶ．

次に包摂関係に基づく同値なパターンについて以下の例を用いて説明する．

$P_1 = gf(X) \leftarrow p(X,Y) \wedge m(Y).$

$P_2 = gf(A) \leftarrow p(A,B) \wedge m(B) \wedge p(A,C) \wedge m(C).$

上の例では $P_1 \supseteq P_2\theta$ となる置換 $\theta = \{A/X, B/Y, C/Y\}$ が存在するので，P_1 は P_2 を包摂する ($p_1 \preceq p_2$) という．同様に $P_2 \preceq P_1$ である．$P_1 \preceq P_2$ かつ $P_2 \preceq P_1$ であるので P_1 と P_2 は同値である ($P_1 \sim P_2$) という．

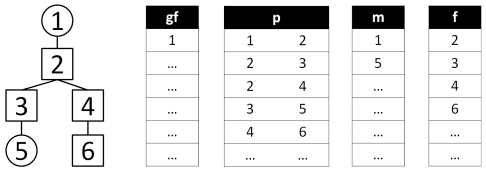

図 1: 家族関係データベース R_{fam}．各関係表は左の家系図に関連するタプルのみを示している．○は男性，□は女性を示し，数字は人物の ID である．

3. Mapix アルゴリズム

Mapix のアイデアは事例の持つ「性質」の組み合わせによるパターンの生成である．アルゴリズム概要を以下に示す．

3.1 Mapix アルゴリズム概要

1) 基本パターン生成　事例の持つ性質 Pr は，以下の条件を満たすリテラルの極小集合である．

1. Pr はただ一つの判定述語をもつ．
2. Pr に含まれる述語の引数は，目標述語の引数から経路述語を経由して判定述語まで繋がっている．

例えば R_{fam} の事例 $gf(1)$ から以下の性質を抽出する．

$Pr_1 = gf(1) \leftarrow p(1,2) \wedge f(2).$

$Pr_2 = gf(1) \leftarrow p(1,2) \wedge p(2,3) \wedge f(3).$

$Pr_3 = gf(1) \leftarrow p(1,2) \wedge p(2,3) \wedge p(3,5) \wedge m(5).$

これらはデータから得られる事実であり，次のように変数化し基本パターンである性質アイテムとする．

$It_1 = gf(A) \leftarrow p(A,B) \wedge f(C).$

$It_2 = gf(A) \leftarrow p(A,B) \wedge p(B,C) \wedge f(C).$

$It_3 = gf(A) \leftarrow p(A,B) \wedge p(B,C) \wedge p(C,D) \wedge m(D).$

2) 頻出性質アイテムセット枚挙　性質アイテムセットとは性質アイテムの独立な組み合わせである．例えば It_2 と It_3 を組み合わせたパターンは次のようになる．

$$\langle It_2, It_3 \rangle = gf(A) \leftarrow p(A,B) \wedge p(B,C) \wedge f(C) \wedge p(A,D) \wedge p(D,E) \wedge p(E,F) \wedge m(F).$$

性質アイテムの独立な組み合わせとは目標述語に含まれる変数以外が異なるような変数化である．

3) 分子アイテム生成　分子アイテムとは性質アイテムの構造に基づく組み合わせである．例えば It_2 と It_3 の構造的組み合わせを考えるとき，それぞれの性質アイテムを生成した性質 Pr_2 と Pr_3 を組み合わせる．

$$Pr_2 \bigcup Pr_3 = gf(1) \leftarrow p(1,2) \wedge p(2,3) \wedge f(3) \wedge p(3,5) \wedge m(5).$$

これを変数化したものを分子アイテムと呼ぶ．

$$It_{2-3} = gf(A) \leftarrow p(A,B) \wedge p(B,C) \wedge f(C) \wedge p(C,D) \wedge m(D).$$

A Consideration on Efficiency of Multi-Relational Pattern Mining Method Based-on Basic Patterns
†Yusuke NAKANO　†Nobuhiro INUZUKA
†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

Algorithm 1 CANDIDATE-GENE
入力: 頻出アイテム集合 $Freq$
k-頻出アイテム集合 F_k
出力: $k+1$-候補アイテム集合 C_{k+1}
1. $C_{k+1} := \phi$
2. Foreach $IS_1 = \langle It_1, \ldots, It_{k-1}, It_k \rangle$,
$IS_2 = \langle It_1, \ldots, It_{k-1}, It'_k \rangle \in F_k (It_k < It'_k)$ do
3. 候補 $Cand := \langle It_1, \ldots, It_{k-1}, It_k, It'_k \rangle$ を生成
4. 以下を満たすとき候補 $Cand$ を C_{k+1} に追加:
5. a) $Cand$ の全ての k-部分集合が F_k に存在
6. b) $Cand$ と同値なものが $Freq \cup C_{k+1}$ に存在しない
7. Return C_{k+1}

Algorithm 2 CANDIDATE-GENE-efficient
入力: 頻出アイテム集合 $Freq$
k-頻出アイテム集合 F_k
出力: $k+1$-候補アイテム集合 C_{k+1}
1. $C_{k+1} := \phi$
2. Foreach $IS_1 = \langle It_1, \ldots, It_{k-1}, It_k \rangle$,
$IS_2 = \langle It_1, \ldots, It_{k-1}, It'_k \rangle \in F_k (It_k < It'_k)$ do
3. If $k = 1, IS_1 \not\preceq IS_2$ かつ $IS_2 \not\preceq IS_1$ Then
4. 候補 $Cand := \langle It_1, It'_1 \rangle$ を C_2 に追加
5. If $k \geq 2$ Then
6. 候補 $Cand := \langle It_1, \ldots, It_{k-1}, It_k, It'_k \rangle$ を生成
7. 以下を満たすとき候補 $Cand$ を C_{k+1} に追加:
8. a) $Cand$ の全ての k-部分集合が F_k に存在
9. Return C_{k+1}

4) 頻出な性質/分子アイテムセット枚挙　性質アイテムと分子アイテムを独立に組み合わせてパターンを生成する．例えば性質アイテム It_1 と分子アイテム It_{2-3} を独立に組み合わせると以下のパターンを得る．

$$\langle It_1, It_{2-3} \rangle = gf(A) \leftarrow p(A,B) \wedge f(B) \wedge p(A,C) \wedge p(C,D) \wedge f(D) \wedge p(D,E) \wedge m(E).$$

4 効率的な候補生成手法の提案

MAPIX は APRIORI[1] と同様の手法でアイテムの頻出な組み合わせを探索する．ILP 手法では生成されたパターンが既に出力されているパターンと同値でない事を検査する必要がある．以下では候補生成のステップについて従来手法で用いられる CANDIDATE-GENE と，本稿で提案する包摂関係に基づく効率化を導入した CANDIDATE-GENE-efficient について述べる．またそれぞれの手法を用いるものを MAPIX-naive，MAPIX-efficient と呼ぶ．

4.1 CANDIDATE-GENE　単純な方法を Algo.1 に示す．これは新たな候補が生成されると既生成パターンの中に同値なパターンがないかをチェックする (6 行)．同値関係のチェックはパターンの双方の包摂関係を調べる必要があるが，包摂関係の計算はコストが高いため候補が生成されるたびにこれを行うのは望ましくない．

4.2 CANDITATE-GENE-efficient

補題 1. アイテム It_1 と It_2 に対して，$It_1 \preceq It_2$ のとき，$\langle It_1, It_2 \rangle$ は It_1 と同値である．

補題 2. 任意のアイテム集合 IS_1 と IS_2 に対して，それらのアイテム集合を構成する任意のアイテム間に包摂関係が存在しないとき，IS_1 と IS_2 は同値ではない．

補題より候補アイテム集合を構成する任意のアイテム間に包摂関係をが存在しなければ，その候補と同値なパターンは生成されていない．これを用いて効率化を行った手法を Algo.2 に示す．Algo.1 では候補生成の度に同値チェックが必要であったが，Algo.2 ではこれが必要なくなる．この手法では 2-候補アイテム集合を生成するときのみ，結合した 1-頻出アイテム集合 2 つの包摂関係を確認し互いに包摂しないときにこれらを結合する (3,4 行)．$k+1$-候補アイテム集合を結合するときは，その全ての k-部分集合が k-頻出アイテム集合に存在するかを確認すればよい (8 行)．k-部分集合で k-頻出アイテム集合に存在しないものがあるということは，その候補アイテム集合が頻出でないか，アイテム間に包摂関係があるという事である．

表 1. 実験結果

	MAPIX-naive	MAPIX-efficient
Bongard dataset (事例数: 320)		
実行時間	4m8s	3m51s
包摂チェック	816,752	143,144
Mutagenesis dataset (事例数: 230)		
実行時間	23m57s	3m7s
包摂チェック	6,152,736	154,362

5 実験

MAPIX-naive と MAPIX-efficient の比較を行った．表.1 より MAPIX-efficient が格段に高速にマイニングできていることがわかる．これは計算コストの高い包摂関係チェック回数が大幅に減少しているためである．また WARMR[3] が Mutagenesis で実行したところ現実的な時間で探索を終了できなかったのに対し，MAPIX-efficient は 3 分程度で探索を終了した．

6 まとめ

本稿では MAPIX の従来の候補生成手法に対して，包摂関係に基づく効率化の仕組みを導入した手法を提案した．これを用いる MAPIX-efficient が従来手法と比べて格段に効率的であることを実験により示した．また ILP 手法の WARMR が処理を終了できなかったデータに対しても MAPIX-efficient は高速に探索を終了できることを示した．しかし，パターンの完全探索を目指す WARMR と比べると，基本パターンに基づく MAPIX は探索空間に制限を受けている．MAPIX の効率性を備え WARMR の探索空間に迫るアルゴリズムの考案が今後の課題として挙げられる．

参考文献

[1] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, VLDB, pp.487-499, 1944.
[2] Y. Nakano and N. Inuzuka, Multi-Relational Pattern Mining Based-on Combination of Properties with Preserving Their Structure in Examples, ILP2010, pp.181-189, 2010.
[3] L. Dehaspe and L. De Raedt, Mining Association Rules with Multiple Relations, ILP97, pp.125-132, 1997.