4B-3

事例から抽出した特徴に基づく関係型パターンマイニング法と他手法の比較

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．この研究分野は 1994 年の Agrawal らによる，トランザクションデータベースからの頻出アイテム集合発見アルゴリズム Apriori[1] の考案から急速に発展した．従来手法の多くは一つの関係表を対象としているが，一般的なデータベースは複数の関係表から構成される．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．著者ら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] を考案した．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．本稿では我々の手法と Dehaspe らによる代表的 ILP システム WARMR[3] の出力パターンと処理時間について比較を行い提案法の有用性を示す．また提案法をグラフマイニングの領域に適用し猪口らによる AGM[4] におけるパターンとの比較を行う．

2 ILP データマイニング

ILP の枠組みでは関係 rel のタプル $< a_1, ..., a_n >$ 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．

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

また探索空間をコントロールするために述語の引数に入力 (+)/出力 (−) のモード情報を与える．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．ここで $p(+, -)$ のように出力引数を持つ述語を経路述語と呼ぶ．また $m(+), f(+)$ のように出力引数を持たない述語を判定述語と呼ぶ．

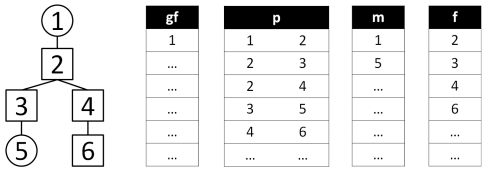


図 1: 家族関係データベース R_{fam}．各関係表は左の家系図に関連するタプルのみを示している．○は男性，□は女性を示し，数字は人物の ID である．

3 提案法: MAPIX

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．アルゴリズム詳細 [2] はページ制約のため省略するが，概要を 4 つのステップに分けて説明する．

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述語の組み合わせである．

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．

$$< it_2, it_3 >= gf(A) \leftarrow p(A, B) \wedge p(B, C) \wedge f(C) \wedge p(A, D) \wedge p(D, E) \wedge p(E, F) \wedge m(F).$$

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．

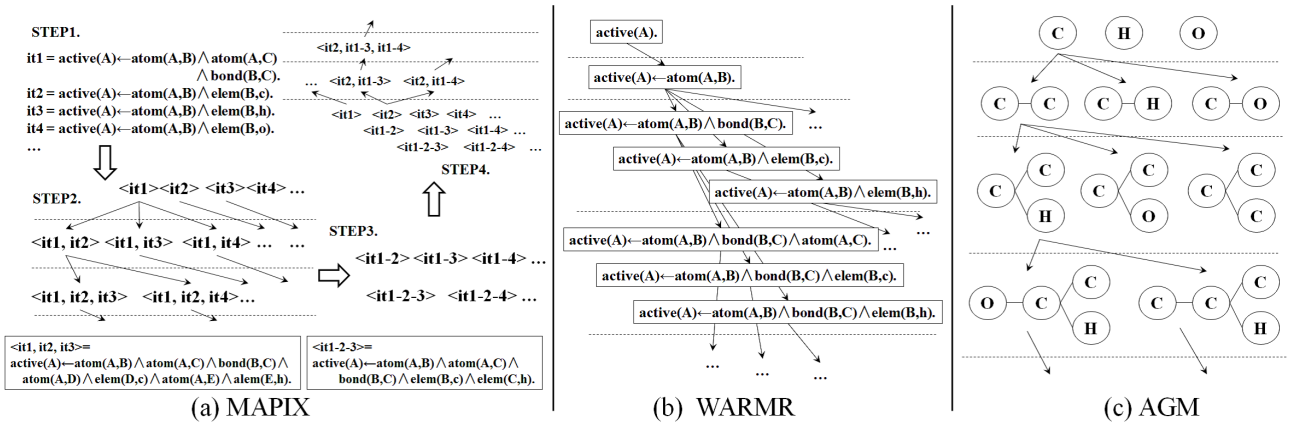
3) 分子アイテム生成　分子アイテムとは性質アイテムの構造に基づく組み合わせである．例えば it_2 と it_3 の構造的組み合わせを考えるとき，それぞれの性質アイテムを生成した性質 pr_2 と pr_3 を組み合わせる．

$$pr_2 \bigcup pr_3 = gf(1) \leftarrow p(1, 2) \wedge p(2, 3) \wedge f(3) \wedge p(3, 5) \wedge m(5).$$

これを変数化したものを分子アイテムと呼ぶ．

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

Multi-Relational Pattern Mining Using Properties Extracted from Examples and Comparison with Other Algorithms
†Yusuke NAKANO　†Nobuhiro INUZUKA
†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology



(a) MAPIX (b) WARMR (c) AGM

図 2: (a) MAPIX, (b) WARMR, (c) AGM のパターン生成手順.

4) 頻出性質/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．

$$< it_1, it_{2-3} >= gf(A) \leftarrow p(A,B) \wedge f(B) \wedge p(A,C) \wedge p(C,D) \wedge f(D) \wedge p(D,E) \wedge m(E).$$

以上の手順で MAPIX は性質アイテムと分子アイテムの頻出な組み合わせを Apriori 同様の手法で枚挙する．

4 他手法との比較

Mutagenesis という変異原性 (遺伝子の突然変異を誘発する能力) をもつ 230 個の化合物に関するデータセットを用いて提案手法 MAPIX と WARMR，AGM の比較を行う．ILP では化合物がもつ原子，結合を以下の関係で表現し入力として与える．

active(*A*) : 化合物 *A* は変異原性をもつ．

atom(+*A*, −*B*) : 化合物 *A* は原子 *B* をもつ．

elem(+*A*, #*B*) : 原子 *A* の原子記号は *B*（定数）である．

bond(+*A*, +*B*) : 原子 *A* と原子 *B* は結合している．

また AGM には各化合物を隣接行列で与える．

4.1 関係型データマイニング

WARMR は Apriori を述語論理に拡張させたアルゴリズムである．図 2(a) および図 2(b) に MAPIX と WARMR のパターン生成手順を示す．WARMR は次のようなパターンを生成する．

a. *active*(*A*) ← *atom*(*A*, *B*).

b. *active*(*A*) ← *atom*(*A*, *B*) ∧ *elem*(*B*,c).

c. *active*(*A*) ← *atom*(*A*, *B*) ∧ *atom*(*A*, *C*) ∧ *bond*(*B*, *C*).

MAPIX は b,c は生成するが a は生成しない．これは探索の基本単位が性質アイテムであり，a は引数が判定述語まで繋がっていないためである．このようなパターンを MAPIX は生成しないため，探索空間は WARMR と比べると制限を受けている．しかしこのデータセットに対して，WARMR が実時間に生成できたパターンは述語 8 個のものであったのに対し，MAPIX は 10 分弱で述語 30 個のパターンを生成した．よって提案法は効率よくマイニングを行う ILP 手法であるといえる．

4.2 グラフマイニング

AGM は Apriori をグラフ理論に拡張させたグラフマイニングアルゴリズムである．図 2(c) の手順で頻出な部分グラフをパターンとして全て生成する (非連結なものも含む)．MAPIX はパターンの探索を性質アイテムの組み合わせと節の包摂関係で制限しているためグラフ構造としては生成できないものがある (例えば C-C という構造は生成できない)．しかし ILP 手法では，述語で定義された知識 (例えば *has_benzen*(*A*, [*B*, *C*, *D*, *E*, *F*, *G*]) : 化合物 *A* は原子 *B*, *C*, *D*, *E*, *F*, *G* で構成されるベンゼン環をもつ) を加えることが可能であり，グラフマイニングの分野では扱えないパターンを扱う事ができる．

5 まとめ

本稿では我々の提案する関係型データマイニングシステム MAPIX を説明し他手法との比較を行った．WARMR との比較では，出力されるパターンに制限は受けるものの効率よくパターンをマイニングできていることを示した．また MAPIX はグラフマイニングを目的としたシステムではないが，グラフの頂点を *node*(*X*)，辺を *edge*(*X*, *Y*) と与えることでこの領域にも適用できることを示した．しかしながらグラフ構造としては出力できないものがあり，データやモード情報の与え方，またグラフマイニングに適したアルゴリズムを考察する必要がある．

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, 2010.

[3] L. Dehaspe and L. De Raedt, Mining Association Rules with Multiple Relations, ILP97, pp.125-132, 1997.

[4] A. Inokuchi, T. Washio and H. Motoda, An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, In Proc. PKDD2000, pp.13-23, LNAI1910, Springer-Verlag, 2000.