平成 21 年度　情報工学科卒業研究概要

基本パターンの出現を保存した融合に基づく関係型パターンマイニング手法

犬塚研究室　知能系
No. 18115115　中野　裕介

1　はじめに

データマイニングとは，大量のデータから隠された知識や新しい規則を発見するプロセスである．その中でも，複数の関係表に跨るパターンを扱う手法を関係型データマイニングといい，代表的なシステムに基本パターンをボトムアップに作成し，その組み合わせからパターンを探索する MaPix[1] がある．本研究では MaPix の基本パターンを組み合わせる方法 (融合)[2] を発展させた手法を提案する．

2　基本パターンの組み合わせによる手法

例えば，家族関係のデータベースにおいて事例 grandfather(koji) にみられる述語の組として以下のようなものがあるとする．

$$\mathsf{grandfather(koji)} \leftarrow \mathsf{parent(koji,\ yozo)} \wedge \mathsf{parent(yozo,\ kyoichi)} \wedge \mathsf{male(kyoichi)}.$$

これは「koji が kyoichi という孫息子を持つ」という事実であり，このように事例が表す述語の組の中でも意味のある述語の組を基本パターンとなる**性質**と考える．この性質の概念を用いて MaPix のアルゴリズムを簡単に説明する．

1. いくつかの事例をサンプリングする
2. サンプリングした事例から性質を全て抽出する
3. 性質を変数化しパターンとなる**性質アイテム**にする
4. 性質アイテムの頻出な組み合わせを枚挙する

また性質アイテムを組み合わせる際は，ヘッドの項以外は共通する変数が無いように変数化する．

MaPix の問題点　以下のような事例について考える．この事例から得られる性質は $p_1,\ p_2$ がある．

$$\mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{m(b)} \wedge \mathsf{p(b,\ c)} \wedge \mathsf{m(c)}.$$
$$p_1 = \mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{m(b)}.$$
$$p_2 = \mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{p(b,\ c)} \wedge \mathsf{m(c)}.$$

これらを変数化すると以下のような性質アイテム $i_1,\ i_2$ が生成される．

$$i_1 = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{m}(B).$$
$$i_2 = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{p}(B,\ C) \wedge \mathsf{m}(C).$$

ここで，i_1 は「息子をもつ」という意味であり，i_2 は「孫息子を持つ」という意味である．これらを組み合わせたアイテムセットは次のようになる．

$$\{i_1,\ i_2\} = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{m}(B) \wedge \mathsf{p}(A,\ C) \wedge \mathsf{p}(C,\ D) \wedge \mathsf{m}(D).$$

$\{i_1,\ i_2\}$ は「息子をもち，かつ孫息子を持つ」という意味であるが，本来事例に現れる「息子をもち，その息子の息子である孫息子を持つ」という意味を成さない．[1] ではこのような MaPix では出力出来なかったパターンを出力する方法として**性質の融合**を用いる．これは，変数化する前の性質同士を組み合わせてから変数化を行うもので，これにより生成されたパターンを**分子アイテム**とよぶ．上の例では $i_1,\ i_2$ の融合を考えるとき変数化する前の性質 $p_1,\ p_2$ をまず組み合わせる．

$$p_1 \wedge p_2 = \mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{m(b)} \wedge \mathsf{p(b,\ c)} \wedge \mathsf{m(c)}.$$

そして，これを変数化し次の分子アイテムを得る．これにより本来事例に現れる意味を表すパターンとなる．

$$i_1 - i_2 = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{m}(B) \wedge \mathsf{p}(B,\ C) \wedge \mathsf{m}(C).$$

3　提案手法

基本パターンの出現を保存した融合　次のような事例について考える．また，この事例から得られる性質 $p_1,\ p_2,\ p_3$ を示す．

$$\mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{p(b,\ c)} \wedge \mathsf{m(c)} \wedge \mathsf{p(b,\ d)} \wedge \mathsf{m(d)} \wedge \mathsf{p(d,\ e)} \wedge \mathsf{m(e)}.$$
$$p_1 = \mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{p(b,\ c)} \wedge \mathsf{m(c)}.$$
$$p_2 = \mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{p(b,\ d)} \wedge \mathsf{m(d)}.$$
$$p_3 = \mathsf{g(a)} \leftarrow \mathsf{p(a,\ b)} \wedge \mathsf{p(b,\ d)} \wedge \mathsf{p(d,\ e)} \wedge \mathsf{m(e)}.$$

これらの性質を変数化すると以下の性質アイテム $i_1,\ i_2$ が得られる．

$$i_1 = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{p}(B,\ C) \wedge \mathsf{m}(C).$$
$$i_2 = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{p}(B,\ C) \wedge \mathsf{p}(C,\ D) \wedge \mathsf{m}(D).$$

さて，i_1 と i_2 の融合を考える際，変数化する前の性質同士の組み合わせを考えるが，この場合 $p_1,\ p_3$ と $p_2,\ p_3$ の 2 通りの組み合わせがある．それぞれの組み合わせを変数化したパターンが異なる場合は，それぞれを分子アイテムとして出力する．

$$i_1 - i_2(1) = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{p}(B,\ C) \wedge \mathsf{m}(C) \wedge \mathsf{p}(B,\ D) \wedge \mathsf{p}(D,\ E) \wedge \mathsf{m}(E).$$
$$i_1 - i_2(2) = \mathsf{g}(A) \leftarrow \mathsf{p}(A,\ B) \wedge \mathsf{p}(B,\ C) \wedge \mathsf{m}(C) \wedge \mathsf{p}(C,\ D) \wedge \mathsf{m}(D).$$

アルゴリズム　融合の手順を見直し，単純なアルゴリズムを提案した．

1. MaPix を用いて頻出な性質アイテムセットを枚挙
2. 頻出な性質アイテムセットから分子アイテムを生成
3. 性質アイテムと分子アイテムを用いて再度 MaPix を実行

4　実験結果

2 つのデータを用いて実験を行った．1 つは図形に関するデータ Bongard であり，これはデータの性質のため複数のパターンが生成されない．アルゴリズムの検証と改良による時間短縮を確認した (表 1)．一方は，5 つの事例からなる人工データであり，多様なパターンが得られることを確認した (表 2)．

表 1: Bongard を用いた実験の実行時間 (s)

使用事例数	5	10	20	30	392
[2] による手法	6.7	13.8	37.6	-	-
提案手法	4.9	8.7	17.0	25.5	6191.7

表 2: 実験用データ用いた実験のパターン数

使用事例数	5
[2] による手法	1943 (30.4s)
提案手法	7307 (82.9s)

5　まとめ

本研究では [2] の手法を発展させた手法を提案した．この手法により，従来の手法と比べて処理時間がが改善され，出力されるパターン数が増えた．しかし，これは実験用データ上での結果であり実データでは使用する事例を増やすと融合を行う際に膨大な処理時間を要してしまった．今後の課題として，より効率的な融合の仕方の考案，またこの手法により出力されているパターンに冗長なものがないことを証明する必要がある．

参考文献

[1] J.Motoyama, S.Urasawa, T.Nakano and N.Inuzuka: “A Mining Items Extracted from Sampled Examples”, ILP2006, Revised Selected Papers, S.Muggelrton et al.(eds.), Lecture Notes in Computer Science, vol. 4455, p.p. 335-350, 2007.

[2] 浦沢　真平，中野　智文，犬塚　信博: “関係的属性の下方閉包性を利用したマイニングアルゴリズム”，第 6 回 人工知能学会 データマイニングと統計数理研究会 (SIG-DMSM)，2008.