平成 19 年度　修士論文概要

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に関する研究

和田・犬塚 研究室　浦澤 真平
No. 18417523　Shinpei Urazawa

1 はじめに

複数の関係表で表されたデータベースから知識や規則を取り出すプロセスを関係データマイニングという．多くの場合データマイニングは 1 つの関係表や属性値表でマイニングが行われるが，一般的なデータは複数の関係表に渡るデータベースで表される．よって，関係データマイニングをおこなう必要がある．

関係データマイニングをおこなう手法として，**帰納論理プログラミング** (Inductive Logic Programming : ILP) が注目される．これは可読性の高い論理的記述によってパターンを出力するもので，関係データマイニングの有力な手法と考えられている．

ILP の枠組みにおけるデータマイニング (以下，ILP データマイニング) の手法として **WARMR** がよく知られている [1]．WARMR は単純なパターンから複雑なものへとトップダウンに発見していくアルゴリズムで，頻出なパターンを枚挙する．しかし，生成される候補が膨大になり計算時間が大きい．

そこで，元山らはボトムアップにパターンの抽出を行う MAPIX(Mininga Algorithm by Property Item eXtraction) を提案した [2]．これは事例に見られるパターンの中でも彼らが**性質**と呼ぶものとその組合せに制限し，興味深いパターンを導出するものである．

MAPIX を発展させた EQUIVPIX[3] は，ボトムアップ手法と性質を単位にすることを維持しつつ，広範囲なパターンを導出することに成功した．しかし，その方法はアドホックであり，効率の点でも改善の余地がある．

本研究では EQUIVPIX で開発された方法を発展させ，EQUIVPIX では得られなかったルールを，相関ルールマイニングでも用いられる下方閉包性を利用し，効率的に導出する．また，この方法が冗長なパターンを導出していないことを証明した．

2 ボトムアップマイニング手法

例えば，家族関係 $\mathbf{R}_{\text{fam}}$ で関係 grandfather に注目してマイニングするとしよう．その内の 1 事例 (タプル) が grandfather(koji) であるとき，ここに現れ koji に関連する他の関係 parent, male, female のタプルが見つかるとき，これを次のように表す

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

これは「koji が kyoichi という孫息子をもつ」という事実である．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**性質**と考える．この性質を使用して，意味のあるパターンに限定してマイニングを行うことが MAPIX のアイディアである．

性質　性質を定義するために，モードを導入する．モードとは，述語の各引数が入力引数であるのか出力引数であるのかを表す情報で，**入力モード** $\oplus$ と**出力モード** $\ominus$ がある．例えば，家族関係 $\mathbf{R}_{\text{fam}}$ に現れる述語のモードは，$\mathsf{parent}(\oplus, \ominus)$，$\mathsf{male}(\oplus)$，$\mathsf{female}(\oplus)$ である．

モードについて注目すると，すべての引数が入力モードである**判定述語**と入力・出力の両モードを持つ**経路述語**という 2 つのクラスに分けられる [2]．

例えば，家族関係 $\mathbf{R}_{\text{fam}}$ では male，female は判定述語，parent は経路述語である．また，各述語で構成されたリテラルをそれぞれ**判定リテラル**，**経路リテラル**と呼ぶ．

判定述語，経路述語の考え方を使って，先ほどの述語の組に注目すると，次のことがいえる．(1) 事実を表す「判定リテラル」が一つだけある．(2)「経路リテラル」によって，引数が事例から判定リテラルまで鎖上に繋がっている．

MAPIX はこのような性質をいくつかの事例から取り出し，得られた性質を Apriori アルゴリズム [4] を使って頻出なパターンを取り出す手法である．

性質の融合によるパターン　MAPIX では出力できなパターンが存在する．例えば家族関係 $\mathbf{R}_{\text{fam}}$ のデータにおいて，事例 grandfather(koji) に以下のタプル組がみつかったとする．

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

この事例は，次の 2 つの性質をもつ．

$$p_1 : \mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A, B) \wedge \mathsf{male}(B). \quad \text{(息子をもつ)}$$

$$p_2 : \mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A, B) \wedge \mathsf{parent}(B, C) \wedge \mathsf{male}(C). \quad \text{(孫息子をもつ)}$$

事例 koji は p_1，p_2 をどちらも満たしているので，$p_1 \wedge p_2$ を満たす．$p_1 \wedge p_2$ は「息子をもつ ∧ 孫息子をもつ」という意味である．しかし，事例 koji には「息子と孫息子がおり，孫息子はその息子の息子である」という事実もあるがこれは p_1 と p_2 の組合せでは表せない．

これらを考慮すると，性質の組合せを得るための 2 種類の操作を考えられる．第 1 は単純な連言である，$\langle p_1, p_2 \rangle$ と書く．

$$\langle p_1, p_2 \rangle : \mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A, B) \wedge \mathsf{male}(B) \wedge \mathsf{parent}(A, C) \wedge \mathsf{parent}(C, D) \wedge \mathsf{male}(D).$$

第 2 は p_1 と p_2 がもともとのタプルの項の値を反映して，同じ項をもつ部分は同じ変数となるように組み合わせるもので $\langle p_1 - p_2 \rangle$ と書く．

$$\langle p_1 - p_2 \rangle : \mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A, B) \wedge \mathsf{male}(B) \wedge \mathsf{parent}(B, C) \wedge \mathsf{male}(C).$$

これを性質の融合といい，できたパターンを長さ 2 の分子アイテムと呼ぶ．

EQUIVPIX では**バインドアイテム**という，アイテムとアイテムが融合していることを表すアイテムを定義し，性質アイテムとバインドアイテムの組合せを探索する手法である．

しかし，この EQUIVPIX ではバインドアイテムを用いるために Apriori アルゴリズムで用いられている効率化ができないことや，バインドアイテムでは表現できないパターンが存在するなどの欠点がある．

3 提案手法

本研究では分子アイテムとその組合せを全て枚挙するアルゴリズムを提案する．その際には次に述べる分子アイテムに関する下方閉包性を用いて，効率的な分子アイテムの生成を行う．そして生成した分子アイテムを用いて MaPix のように分子アイテムの組合せを探索する．

分子アイテムの持つ性質　Apriori アルゴリズムではアイテムセットの支持度に関する下方閉包性，すなわち，頻出なアイテム集合の部分集合もまた頻出であるという性質を利用することで効率的な計算を可能にしていた．

分子アイテムの支持度も同様に下方閉包性を満たす．すなわち a, b を分子アイテムとすると次の式を満たす．

$$a \subseteq b \Rightarrow \mathrm{support}(a) \geq \mathrm{support}(b) \qquad (1)$$

また，あるアイテム集合を融合させた場合と組み合わせた場合では組み合わせたパターンの方が支持度が高くなる．

$$\mathrm{support}(\langle i_1, \cdots, i_n \rangle) \geq \mathrm{support}(\langle i_1 - \cdots - \imath_n \rangle) \quad (2)$$

式 (1) より，分子アイテムも Apriori 同様に level-wise に生成する．また k-分子アイテムを計算する際には，k-アイテムセットを先に計算しておき，式 (2) を用いて候補を減らすことで効率的な計算が可能になる．

アルゴリズム　前節で述べた分子アイテムの性質を用いるために，次図で示す順序で分子アイテムの生成とその組合せの探索を行う．図中では，探索順序と共に，n 個以下のアイテムを融合した分子アイテムを k 個組み合わせたアイテムセットであることを n-k と表記し，その例を示した．図中の矢印は，各パターングループの生成に利用するグループであり，その際に下方閉包性が利用できる．本手法では下方閉包性以外に，同値パターンを同値性を毎回判定せずに扱うための仕組みを導入した．

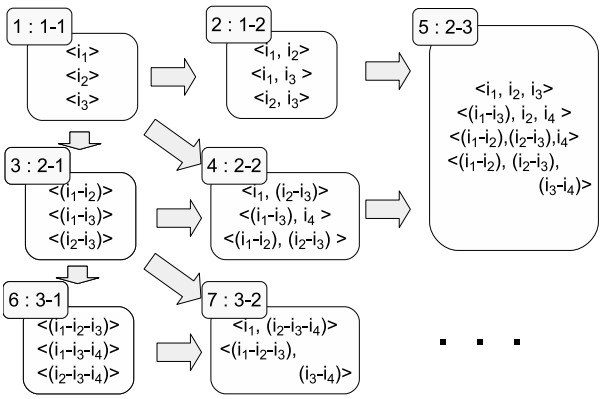

図 2: 探索順序

4 実験

実験データとして Bongard という図形に関するデータを使用し，WARMR，MaPix，EquivPix と提案手法で得られるパターンの数と処理時間を比較する．最低支持度を 5%とし，性質の抽出に使用する事例を変えて比較する．

処理時間とパターン数について表 1，2 に示す．

表 1：実行時間 (s)

使用事例数	5	10	20	30	全事例
WARMR	1098.5				
MaPix	3.6	7.1	13.6	19.2	245.6
EquivPix	5.7	12.2	30.0	44.3	280.0
提案手法	9.4	20.6	40.5	146.1	$> 10^4$

表 2：パターン数

	5	10	20	30	全事例
WARMR	5480(同値類:782)				
MaPix	145	150	160	160	160
EquivPix	270	424	542	599	625
提案手法	446	621	764	752	-

提案手法の実行時間は，表 1 に示した範囲では MaPix，EquivPix におとるものの WARMR よりは少ない時間で処理できている．しかし使用事例数を 40 以上にしたときには，4 時間以上かかり WARMR よりも十分に遅いため実験を中止した．

しかし，表 2 に示した通り事例数が 30 までででも EquivPix よりも多く WARMR にせまる数のルールを出力できている．

5 おわりに

本研究では ILP データマイニングの手法である EquivPix を発展させた手法を提案した．この手法は EquivPix よりも多彩なパターンを出力することができ，その際には同値なパターンを出力していないことを証明した．

今後の課題としては，性質の抽出に使用する事例数を増やしたときに処理時間が爆発的に増えるという問題を解決する必要がある．

参考文献

1. L. Dehaspe, H. Toivonen :“Discovery of Relational Association Rules”, in Relational Data Mining, pp. 189–212, Springer, 2001.

2. 元山純一，中野智文，犬塚信博：“関係的相関ルール導出のための事例の性質の抽出”，FIT2005，pp. 5–8，2005

3. J. Motoyama, S. Urazawa, T. Nakano and N. Inuzuka: A Mining Algorithm Using Property Items Extracted from Sampled Examples. Proc. ILP 2006, Revised Selected Papers, pp.335-350, 2007.

4. R. Agrawal, R. Srikant : “Fast Algorithms for Mining Association Rules”, Proc. VLDB, pp. 487-499, Morgan Kaufmann, 1994.