2A3-3

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

Multi-Relational Pattern Mining Based-on Combination of Properties with Preserving Their Structure in Examples

中野　裕介[*1]　　犬塚　信博[*1]
Yusuke Nakano　Nobuhiro Inuzuka

[*1]名古屋工業大学　大学院　工学研究科　情報工学専攻
Depertment of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

We propose a new algorithm for the problem of multi-relational pattern mining through the problem established in Warmr. In order to overcome the combinatorial problem of large pattern space, another algorithm Mapix restricts patterns into combination of basic patterns, called properties. A property is defined as a set of literals appeared in examples and is an extended form of the attribute-value form. Mapix enumerates patterns made from conjunction of the properties. Although the range of patterns is clear and Mapix enumerates them efficiently, a large part of patterns are out of the range. Advantage of Mapix is to make patterns from pattern fragments occured in examples. Many patterns which are not appeared in examples are not tested. The proposing algorithm keeps this advantages and extends the way of combination of properties. The algorithm adopts a way of combination it combines properties as they appeared in examples.

1. はじめに

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をデータマイニングとよび，その中でも複数の関係表に跨るようなパターンを扱うものを関係型データマイニング (Multi-Relational Data Mining: MRDM) とよぶ．従来のデータマイニングの多くは 1 つの関係表を用いたマイニングであるが，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ため MRDM を行う必要がある．

MRDM は帰納論理プログラミング (Inductive Logic Programming: ILP) の枠組みで行われてきた．ILP は述語倫理を用いた記述により，豊かな表現力と高い可読性を持っており，MRDM の有力な手法と考えられている．

MRDM の代表的な手法として Warmr[2] がある．これは，単純なパターンから複雑なパターンへとトップダウンに頻出なパターンを探索するアルゴリズムである．しかし，生成される候補が膨大にあり計算時間が大きい．

これに対し，元山らはボトムアップに基本パターンを生成し，その組み合わせによりパターンを探索する Mapix(Mining Algorithm by Property Item eXtraction)[4] を考案した．これは，引数のモードやタイプ等で制限を行いながら，事例に関する事実を取り出し，それらをアイテムとして興味深いルールを導出するものである．これにより探索の高速化に成功したが，出力されるパターンに制限を受けてしまった．

そこで，Mapix を拡張させた Equivpix[3] は基本パターンを構造に基づいて組み合わせるオペレーションにより探索の範囲を広げた．しかし，そこで用いられているオペレータは全ての組み合わせを見ておらず改善の余地がある．

本研究では Equivpix を発展させたアルゴリズムを提案する．提案手法では Mapix を 2 回用いる．1 回目は従来と同様に頻出な基本パターンの組み合わせを枚挙する．2 回目の Mapix を行う前に 1 回目の出力結果を用いて構造的な組み合わせを行い，その出力結果を用いて再度 Mapix を実行する．

連絡先: 名古屋工業大学 大学院 工学研究科 情報工学専攻
〒 466-8555 愛知県名古屋市御器所町
E-mail: nakano18115115@gmail.com

2. 準備

ILP は述語論理の一般的な定義に基づく．MRDM では，それぞれの関係表を述語論理の形式で表現しパターンをマイニングする．例として家族関係データベース$R_{\mathbf{fam}}$について考える (図 1)．$R_{\mathbf{fam}}$は 4 つの関係表をもち，それぞれを次のように述語論理の形式で表現する．$\mathsf{parent}(x,y)$ は x が y の親であることを，$\mathsf{male}(x)$/$\mathsf{female}(x)$ は x が男性/女性であることを，$\mathsf{grandfather}(x)$ は x が祖父であることを意味する．01 から 24 は人物を表すインデックスである．以下では省略して gf，p，m，f を用いる．まず Mapix におけるパターンを定義する．ここで $target$ は特徴を取り出したい事例の述語で構成されたリテラルであり目標事例 (目標述語) とよぶ．これは Warmr における key の概念に相当する．

定義 1 (パターン) パターンは次のような節である．

$$P = target \leftarrow char_1 \wedge \cdots \wedge char_n$$

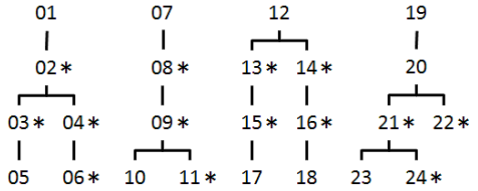
$e = head(P)\rho$ となる最汎単一化代入 ρ を使って，$\mathbf{R} \models body(P)\rho\theta$ となる代入 θ が必ず存在するとき，事例 e がデータベース $\mathbf{R}$ 上でパターン P を満たすという．

定義 2 (頻出パターン) パターン P の支持度とは，P を満たす事例の割合であり $\mathbf{supp}(P)$ と表記する．$\mathbf{supp}(P) \geq \mathrm{sup}_{\min}$ のときパターン P は頻出であるという．ただし，$\mathrm{sup}_{\min}$ はユーザにより与えられる最低支持度である．

例えば，次の式は$R_{\mathbf{fam}}$におけるパターンである．

$$P = \mathsf{gf}(A) \leftarrow \mathsf{m}(A) \wedge \mathsf{p}(A,B) \wedge \mathsf{f}(B).$$

このパターンを事例 $e = \mathsf{gf}(01)$ が満たしているかどうか調べるとき，$e = head(P)\rho$ となるような代入 $\rho = \{A/01\}$ を考える．$body(P)\rho = \mathsf{m}(01) \wedge \mathsf{p}(01,B) \wedge \mathsf{f}(B)$ の中の全ての述語が$R_{\mathbf{fam}}$に現れるような代入 $\theta = \{B/01\}$ が存在するので，事例 $e = \mathsf{gf}(01)$ はパターン P を満たす．同様に事例 $\mathsf{gf}(07)$，$\mathsf{gf}(12)$，$\mathsf{gf}(20)$ も P を満たすが，$\mathsf{gf}(19)$ はこれを満たさない．つまり P の支持度は $\mathbf{supp}(P) = 4/5 = 80\%$ となり，最低支持度が 80%以下のとき P は頻出パターンとなる．



grandfather	parent		male	female
01	01	02	01	02
07	02	03	05	03
12	02	04	07	04
19	03	05	10	06
20	...	...	...	...

図 1: 4 つの関係表 grandfather，parent，male，female からなる家族関係のデータベース$R_{\mathbf{fam}}$．$*$ がついているものは女性を，それ以外は男性を表す．

MRDM では，効率よく探索を行うために論理的に意味の同じものを出力しないようにすることがある．そこで，同値なパターンについて述べるため包摂関係を定義する．

定義 3 (包摂関係，同値) 節 C，D について $C\theta \subseteq D$ をみたす代入 θ が存在するとき，節 C は D を包摂するといい $C \succeq D$ と表す．また，$C \succeq D$ かつ $D \succeq C$ が成り立つとき，節 C と D は包摂に関して同値であるといい $C \sim D$ と表す．

また，ILP におけるアルゴリズムの多くはパターンを制限するために述語の引数にモードを利用している．これは各引数が入力引数であるか出力引数であるかを表す情報で，前者を入力モード，後者を出力モードという．$R_{\mathbf{fam}}$に現れる述語にはそれぞれ $\mathsf{p}(+,-)$，$\mathsf{m}(+)$，$\mathsf{f}(+)$ のようにモードが与えられている．ただし $+/-$ は入力/出力を意味する．

さて，述語のモードについて注目すると，述語は 2 つのクラスに分けることができる．例えば $\mathsf{p}(+,-)$ のように少なくとも 1 つの $\langle - \rangle$-引数を持つ述語を経路述語とよび，これは入力の項から出力の項へ導く機能をもつ．$\mathsf{m}(+)$ と $\mathsf{f}(+)$ のように $\langle - \rangle$-引数を持たない述語を判定述語とよび，その項が持つ特徴を表す機能をもつ．また，経路/判定述語で構成されたリテラルを経路/判定リテラルとよぶ．

以上で説明した概念を用いて Mapix における基本パターンである性質と，それを変数化させた性質アイテムを定義する．

定義 4 (性質) データベース $\mathbf{R}$ について事例 e に関する性質 pr は，以下の条件を満たす $\mathbf{R}$ 中のリテラルの極小集合である．

1. pr は，必ずただ一つの判定リテラルを含む．
2. pr は，全てのリテラル $l (l \in \mathsf{pr})$ の入力引数に現れる項が，そのリテラルより前のリテラルの引数に現れるように順序づけを与えることができる．

定義 5 (変数化) ある基礎節 α について，以下の条件を満たすような節 β は α の変数化である．

1. β は基礎項を持つようなリテラルを含まない．
2. 以下の条件を満たす代入 $\theta = \{v_1/t_1, \cdots, v_n/t_n\}$ が存在する．
   (a) $\alpha = \beta\theta$ (b) $t_1, \cdots, t_n$ は，すべて異なる項である．

定義 6 (性質アイテム) 基礎リテラルの集合 $L = \{l_1, \cdots, l_m\}$ と事例 e について，$e \leftarrow l_1 \wedge \cdots \wedge l_m$ を変数化したものを $\mathbf{var}(e \leftarrow L)$ と表記する．L が e の性質であるとき，$\mathbf{var}(e \leftarrow L)$ を性質アイテムと呼ぶ．

表 1: e=gf(01) の性質と性質アイテム

性質 (*properties*)
pr0= {m(01)}
pr1= {p(01,02),f(02)}
pr2= {p(01,02),p(02,03),f(03)} ({p(01,02),p(02,04),f(04)})
pr3= {p(01,02),p(02,03),p(03,05),m05}
pr4= {p(01,02),p(02,04),p(04,06),f06}
性質アイテム (*property items*)
it0 = $\mathsf{gf}(A) \leftarrow \mathsf{m}(A)$.
it1 = $\mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{f}(B)$.
it2 = $\mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(B,C) \wedge \mathsf{f}(C)$.
it3 = $\mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(B,C) \wedge (C,D) \wedge \mathsf{m}(D)$.
it4 = $\mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(B,C) \wedge (C,D) \wedge \mathsf{f}(D)$.

$R_{\mathbf{fam}}$において，$\mathsf{pr} = \{\mathsf{p}(01,02), \mathsf{p}(02,03), \mathsf{f}(03)\}$ は事例 $e = \mathsf{gf}(01)$ の性質である．これから生成される性質アイテムは it=$\mathbf{var}(e \leftarrow \mathsf{pr})$=$\mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(B,C) \wedge \mathsf{f}(C)$ となる．表 1 に事例 gf(01) から抽出される性質と性質アイテムを全て示す．pr2 のように同値な性質が存在する場合は同値な性質アイテムを生成しないようにしてパターンの重複を避ける．

以下に Mapix アルゴリズムの概要を示す:

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. 性質を変数化し性質アイテムを生成する
4. アプリオリアルゴリズム [1] と同様の手法でアイテムの頻出な組み合わせを枚挙する

事例 gf(01) から生成される性質アイテムについて，最低支持度を 60%とするとこれらの性質アイテムは全て頻出である．しかし，事例 gf(20) から生成される性質アイテム $\mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(B,C) \wedge \mathsf{m}(C)$ は頻出ではない．

性質アイテムを組み合わせるとき，Mapix では連言で性質アイテムを結合している．例えば「孫娘をもつ」という意味を表す it2 と「ひ孫娘をもつ」という意味を表す it4 を組み合わせたパターンは以下のようになる．

$$\langle \mathsf{it2}, \mathsf{it4} \rangle = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(B,C) \wedge \mathsf{f}(C) \wedge \mathsf{p}(A,D) \wedge \mathsf{p}(D,E) \wedge (E,F) \wedge \mathsf{f}(F).$$

これは「孫娘とひ孫娘をもつ」という意味である．このように Mapix では，それぞれの性質アイテムについて目標述語に含まれる変数以外を他の性質アイテムの変数と共通しないように変数化し性質アイテム同士は独立であるようにしている．つまり，it2 と it4 を組み合わせても実際に事例に現れる「娘と孫娘をもつ子供をもつ」や「娘を持つ孫娘をもつ」という意味を表すパターンは生成されない．

3. 提案手法

Mapix では性質アイテム同士を独立に組み合わせていたため生成できないパターンがあった．そこで，性質アイテムの構造的な組み合わせを用いて Mapix で生成されないパターンを生成する手法を提案する．まず，性質アイテムの事例への出現の仕方を保存するために性質アイテムの影を定義する．

定義 7 データベース $\mathbf{R}$ と性質アイテム it について，以下のような集合 $\mathbf{shadow}(\mathsf{it}, \mathbf{R})$ を性質アイテム it の影という．

$$\mathbf{shadow}(\mathsf{it}, \mathbf{R}) = \{\ `e \leftarrow L' \in T \times 2^{\mathbf{R}} \mid L \text{ は事例 } e \text{ に関する性質であり } \mathbf{var}(e \leftarrow L) \sim \mathsf{it}\}$$

ただし，T は目標述語の事例集合であり，2^X は集合 X のベキ集合である．

次に構造的な組み合わせの方法である性質の融合と，それにより生成される分子アイテムを定義する．

定義 8 (性質の融合) データベース $\mathbf{R}$ について，事例 e に関する変数化をしていない性質の集合を P とする．$\{p_1,\cdots,p_n\}\subseteq P$ について，$L_P = p_1\cup\cdots\cup p_n$ とし，リテラル集合 $L_P=\{\ell_1,\cdots,\ell_k\}$ から節 $e\leftarrow\ell_1\wedge\cdots\wedge\ell_k$ を生成し変数化することを，性質 $p_1,\cdots,p_n$ の融合という．

定義 9 性質アイテムセット $\langle \mathsf{it}_{i_1},\cdots,\mathsf{it}_{i_n}\rangle$ に対し，影の組み合わせ，

$$\langle e\leftarrow \mathsf{pr}_{i_1},\cdots,e\leftarrow \mathsf{pr}_{i_n}\rangle \in \mathbf{shadow}(\mathsf{it}_{i_1},\mathbf{R})\times\cdots\times\mathbf{shadow}(\mathsf{it}_{i_n},\mathbf{R})$$
$$s.t.\ \textstyle\bigcap_{j=1,\ldots,n}(\mathbf{terms}(\mathsf{pr}_{i_j})-\mathbf{terms}(e))\neq\emptyset$$

が存在するとき，この性質アイテムセットは融合可能であるという．ただし，$\mathbf{terms}(p)$ は p の中の全ての項の集合である．

定義 10 (分子アイテム) 性質アイテムセット $\mathsf{is}=\langle \mathsf{it}_{i_1},\cdots,\mathsf{it}_{i_n}\rangle$ について融合可能な影の組み合わせ $\langle e\leftarrow \mathsf{pr}_{i_1},\cdots,e\leftarrow \mathsf{pr}_{i_n}\rangle$ が存在するとき，

$$\mathbf{var}(e\leftarrow\textstyle\bigcup_{j=1,\ldots,n}\mathsf{pr}_{i_j})$$

を is から生成された分子アイテムとよぶ．

例えば，性質アイテム it2 と it4 の影は次のようになる．

$\mathbf{shadow}(\mathsf{it2},R_{\mathbf{fam}})=\{$
$\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,03),\mathsf{f}(03)\}$,
$\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,04),\mathsf{f}(04)\}$,
$\mathsf{gf}(07)\leftarrow\{\mathsf{p}(07,08),\mathsf{p}(08,09),\mathsf{f}(09)\}$,
$\mathsf{gf}(12)\leftarrow\{\mathsf{p}(12,13),\mathsf{p}(13,15),\mathsf{f}(15)\}$,
$\mathsf{gf}(12)\leftarrow\{\mathsf{p}(12,14),\mathsf{p}(14,16),\mathsf{f}(16)\}$,
$\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,21),\mathsf{f}(21)\}$,
$\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,22),\mathsf{f}(22)\}$,
$\mathsf{gf}(20)\leftarrow\{\mathsf{p}(20,21),\mathsf{p}(21,24),\mathsf{f}(24)\}$ $\}$

$\mathbf{shadow}(\mathsf{it4},R_{\mathbf{fam}})=\{$
$\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,04),\mathsf{p}(04,06),\mathsf{f}(06)\}$,
$\mathsf{gf}(07)\leftarrow\{\mathsf{p}(07,08),\mathsf{p}(08,09),\mathsf{p}(09,11),\mathsf{f}(11)\}$,
$\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,21),\mathsf{p}(21,24),\mathsf{f}(24)\}$ $\}$

性質アイテムセット $\langle\mathsf{it2},\mathsf{it4}\rangle$ の融合を考えるとき，融合可能な影の組み合わせの有無をみる．$\langle\mathsf{it2},\mathsf{it4}\rangle$ は以下の 5 つの融合可能な影の組み合わせを持つ．

$\langle\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,03),\mathsf{f}(03)\}$,
$\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,04),\mathsf{p}(04,06),\mathsf{f}(06)\}\rangle$
$\langle\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,04),\mathsf{f}(04)\}$,
$\mathsf{gf}(01)\leftarrow\{\mathsf{p}(01,02),\mathsf{p}(02,04),\mathsf{p}(04,06),\mathsf{f}(06)\}\rangle$
$\langle\mathsf{gf}(07)\leftarrow\{\mathsf{p}(07,08),\mathsf{p}(08,09),\mathsf{f}(09)\}$,
$\mathsf{gf}(07)\leftarrow\{\mathsf{p}(07,08),\mathsf{p}(08,09),\mathsf{p}(09,11),\mathsf{f}(11)\}\rangle$
$\langle\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,21),\mathsf{f}(21)\}$,
$\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,21),\mathsf{p}(21,24),\mathsf{f}(24)\}\rangle$
$\langle\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,22),\mathsf{f}(22)\}$,
$\mathsf{gf}(19)\leftarrow\{\mathsf{p}(19,20),\mathsf{p}(20,21),\mathsf{p}(21,24),\mathsf{f}(24)\}\rangle$

これらの組み合わせから融合を行い同値なものを生成しないようにすると，次の 2 つの分子アイテムが生成される．

$\mathsf{it2\text{-}4} = \mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(C,D)\wedge\mathsf{f}(D)$.
$\mathsf{it2\text{-}4'} = \mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(B,D)\wedge\mathsf{p}(D,E)\wedge\mathsf{f}(E)$.

it2-4 は「娘をもつ孫娘をもつ」という意味を表し，it2-4' は「娘と孫娘をもつ子供をもつ」という意味を表す．これらは MAPIX では生成できなかったパターンである．$R_{\mathbf{fam}}$における頻出な性質アイテムと分子アイテムの全てを表 2 に示す．

表 2: 最低支持度が 60%のときの，$R_{\mathbf{fam}}$における頻出な性質アイテムと分子アイテム．

it0	=	$\mathsf{gf}(A)\leftarrow\mathsf{m}(A)$.
it1	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{f}(B)$.
it2	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)$.
it3	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{p}(C,D)\wedge\mathsf{m}(D)$.
it4	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{p}(C,D),\mathsf{f}(D)$.
it1-2	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{f}(B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)$.
it1-3	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{f}(B)\wedge\mathsf{p}(B,C)\wedge\mathsf{p}(C,D)\wedge\mathsf{m}(D)$.
it2-3	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(C,D)\wedge\mathsf{m}(D)$.
it2-3'	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(B,D)\wedge\mathsf{p}(D,E)\wedge\mathsf{m}(E)$.
it2-4	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(C,D)\wedge\mathsf{f}(D)$.
it2-4'	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(B,D)\wedge\mathsf{p}(D,E)\wedge\mathsf{f}(E)$.
it3-4	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{p}(C,D)\wedge\mathsf{f}(D)\wedge\mathsf{p}(B,E)\wedge\mathsf{p}(E,F),\mathsf{m}(F)$.
it1-2-3	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{f}(B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(C,D)\wedge\mathsf{m}(D)$.
it1-2-3'	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{f}(B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(B,D)\wedge\mathsf{p}(D,E)\wedge\mathsf{m}(E)$.
it2-3-4	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{f}(C)\wedge\mathsf{p}(C,D)\wedge\mathsf{f}(D)\wedge\mathsf{p}(B,E)\wedge\mathsf{p}(E,F)\wedge\mathsf{m}(F)$.
it2-3-4'	=	$\mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{p}(B,C)\wedge\mathsf{p}(C,D)\wedge\mathsf{f}(D)\wedge\mathsf{p}(B,E)\wedge\mathsf{f}(E)\wedge\mathsf{p}(E,F)\wedge\mathsf{m}(F)$.

また，性質アイテムと分子アイテムを独立に組み合わせるとあらたなパターンを得ることができる．例えば，it1 と it2-4 を組み合わせたアイテムセットは次のようになる．

$$\langle\mathsf{it1},\mathsf{it2\text{-}4}\rangle = \mathsf{gf}(A)\leftarrow\mathsf{p}(A,B)\wedge\mathsf{f}(B)\wedge\mathsf{p}(A,C)\wedge\mathsf{p}(C,D)\wedge\mathsf{f}(D)\wedge\mathsf{p}(D,E)\wedge\mathsf{f}(E).$$

これは「娘と娘を持つ孫娘をもつ」という意味を表す．

以上の考えを用いて提案手法の概要を次に示す (詳細は表 3):

1. 目標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抽出し性質アイテムを生成する
3. アプリオリ同様のアルゴリズムで性質アイテムの頻出な組み合わせを全て枚挙する
4. 性質アイテムの頻出な組み合わせから分子アイテムを生成する
5. 再度アプリオリと同様のアルゴリズムで性質アイテムと分子アイテムの頻出な組み合わせを全て枚挙する

ステップ 4 では頻出な性質アイテムセットから分子アイテムを生成する．これは性質アイテムセット $\langle\mathsf{it1},...,\mathsf{itN}\rangle$ と分子アイテム it1-...-N の支持度が $\mathbf{supp}(\langle\mathsf{it1},...,\mathsf{itN}\rangle)\leq\mathbf{supp}(\mathsf{it1}\text{-}...\text{-}\mathsf{N})$ の関係にあるためである．この式は性質アイテムセットよりそれから生成される分子アイテムの方が支持度が低くなるということを表している．つまり頻出でない性質アイテムセットから生成される分子アイテムもまた頻出ではないので生成する必要がない．

4. 実験とまとめ

2 つのデータを用いて実験を行った．1 つの実験は$R_{\mathbf{fam}}$を用いたもので提案手法により従来よりも多種のパターンが生

表 3: 提案手法のアルゴリズム.

input $\boldsymbol{R}$: 関係 DB; T : 目標事例; $\sup_{\min}$: 最低支持度 (%);
output *Freq*: 頻出アイテムセットの集合;

1. サンプリングする目標事例を選択 $T' \subseteq T$;
2. *Items* := ø; $\mathcal{P}$:= ø; *Freq*:= ø;
3. **For each** $e \in T'$ **do** $\mathcal{P}$:=$\mathcal{P} \cup \{e \leftarrow \mathsf{pr} \mid \mathsf{pr}$ は e の性質 $\}$;
4. **For each** '$e \leftarrow \mathsf{pr}$' $\in \mathcal{P}$ **do**
5. **If** $\exists I \in$ *Items*, $I \sim$**var**$(e \leftarrow \mathsf{pr})$ **then**
6. $S[I]$:=$S[I] \cup \{e \leftarrow \mathsf{pr}\}$; % $S[I]$=**shadow**$(I,\boldsymbol{R})$
7. **else** I' = **var**$(e \leftarrow \mathsf{pr})$;
8. $S[I']$:=$\{e \leftarrow \mathsf{pr}\}$;
9. *Items*:=*Items* $\cup \{I'\}$;
10. k:=1;
11. $\mathcal{F}_1^1$:=$\{\langle I \rangle \mid I \in$ *Items* かつ **supp**$(I) \geq \sup_{\min}\}$;
12. *Freq*:=$\mathcal{F}_1^1$;
13. **While** $\mathcal{F}_k^1 \neq$ ø **do**
14. $\mathcal{C}_{k+1}$:=CANDIDATE$(\mathcal{F}_k^1, \mathcal{F}_k^1)$;
15. $\mathcal{F}_{k+1}^1$:=$\{IS \in \mathcal{C}_{k+1} \mid$ **supp**$(IS) \geq \sup_{\min}\}$;
16. *Freq*:=*Freq* $\cup \mathcal{F}_{k+1}^1$;
17. k:=k+1;
18. *Comb* := CANDICOMB(*Freq*);
19. k:=1;
20. $\mathcal{F}_1^2$:=$\{\langle I \rangle \mid I \in$ *Comb* かつ **supp**$(I) \geq \sup_{\min}\}$;
21. *Freq*:=*Freq* $\cup \mathcal{F}_1^2$;
22. **While** $\mathcal{F}_k^2 \neq$ ø **do**
23. $\mathcal{C}_{k+1}$:=CANDIDATE$(\mathcal{F}_k^1, \mathcal{F}_k^2)$;
24. $\mathcal{F}_{k+1}^2$:=$\{IS \in \mathcal{C}_{k+1} \mid$ **supp**$(IS) \geq \sup_{\min}\}$;
25. *Freq*:=*Freq* $\cup \mathcal{F}_{k+1}^2$;
26. k:=k+1;
27. **Return** *Freq*;

CANDIDATE$(\mathcal{F}_k^1, \mathcal{F}_k^2)$:
input $\mathcal{F}_k^1, \mathcal{F}_k^2$: あるレベルの頻出アイテムセットの集合;
output $\mathcal{C}_{k+1}$: 次のレベルの候補アイテムセットの集合;

1. $\mathcal{C}_{k+1}$:= ø;
2. **For each** $\langle\langle I_1, \ldots, I_k\rangle, \langle I'_1, \ldots, I'_k\rangle\rangle \in \mathcal{F}_k^2 \times \{\mathcal{F}_k^1 \cup \mathcal{F}_k^2\}$, ただし $I_1 = I'_1, \ldots, I_{k-1} = I'_{k-1}$ かつ $I_k < I'_k$ **do**
3. $\mathcal{C}_{k+1} := \mathcal{C}_{k+1} \cup \{\langle I_1, \ldots, I_{k-1}, I_k, I'_k\rangle\}$;
4. **For each** $IS \in \mathcal{C}_{k+1}$ **do**
5. **If** $k = 1$ かつ ($I \preceq I'$ または $I' \preceq I$), ただし $IS = \langle I, I'\rangle$ **then**
6. $\mathcal{C}_{k+1}$ から IS を取り除く;
7. **For each** $I \in IS$ **do if** $IS - \{I\} \notin \mathcal{F}_k^1 \cup \mathcal{F}_k^2$ **then**
8. $\mathcal{C}_{k+1}$ から IS を取り除く;
9. **Return** $\mathcal{C}_{k+1}$;

CANDICOMB(*Freq*):
input *Freq* : 頻出性質アイテムセットの集合;
output *Comb* : *Freq* から生成される分子アイテムの集合;

1. *Comb* := ø;
2. **For each** $\langle I_1, \ldots, I_n\rangle \in$ *Freq* for $n \geq 2$ **do**
3. **For each** $\langle e_1 \leftarrow \mathsf{pr}_1, \ldots, e_n \leftarrow \mathsf{pr}_n\rangle \in S[I_1] \times \cdots \times S[I_n]$ **do**
4. **If** $e_1 = \ldots = e_n$ かつ $\bigcap_{j=1,\ldots,n}$(**terms**(pr_j) − **terms**$(e_j)) \neq$ ø **then**
5. *Comb* := *Comb* $\cup \{$**var**$(e_1 \leftarrow \bigcup_{j=1,\ldots,n} \mathsf{pr}_j)\}$;
6. **For each** $I \in$ *Comb* **do**
7. **If** $\exists I' \in$ *Comb*, $I' \neq I$ かつ $I' \sim I$ **then**
8. *Comb* から I を取り除く;
9. **Return** *Comb*;

表 4: $R_{\mathbf{fam}}$を用いた実験結果.

	最低支持度	20%	40%	60%	80%
MAPIX	パターン数	55	31	23	11
[4]	時間 (sec.)	0.04	0.01	0.01	0.01
EQUIVPIX	パターン数	441	153	51	15
[3]	時間 (sec.)	6.23	0.45	0.06	0.03
提案手法	パターン数	4601	1063	109	17
	時間 (sec.)	9.55	0.54	0.08	0.01

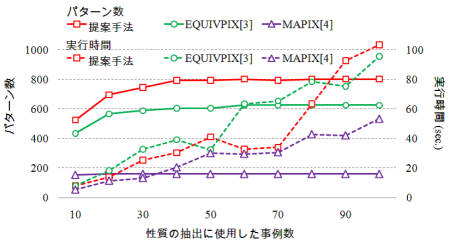


図 2: Bongard を用いた実験結果.

成されることを確認した．表 4 は最低支持度をを変えながら，それぞれのアルゴリズムにより生成されるパターン数と実行時間を示している．これらのパターンの中には同値であるという意味での重複がないことを確認している．提案手法により多種のパターンが生成されていることを確認できる．

もう 1 つの実験は図形に関するデータ Bongard を用いたものである．図 2 ではサンプリングする事例数を変えながら，生成されるパターン数と実行時間を示している．これらはそれぞれ 10 回実行したときの平均である．支持度を計算するには全ての事例を使うが，性質を抽出するときはサンプリングされたいくつかの事例を用いる．提案手法では 80 事例をサンプリングすることで全事例 392 をサンプリングして得られる全パターン (802 個) を生成した．またこれらのパターンに重複がないことを確認している．

本研究では MRDM における MAPIX アルゴリズムを発展させた手法を提案した．提案手法により従来手法よりも多種のパターンを出力することが可能になった．しかし，サンプリング事例のサイズが大きくなると処理時間が爆発的に増えてしまった．実際，Bongard で全事例 392 をサンプリングして実行すると処理時間が 6556 秒かかった．これはサンプリング事例数が増加するに伴い性質アイテムの影の組み合わせが膨大になるためだと思われる．効率的な分子アイテムの生成方法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, 1994.

[2] L. Dehaspe and L. De Raedt. Mining association rules with multiple relations. ILP97, pp.125–132, 1997.

[3] N. Inuzuka, J. Motoyama, S. Urazawa and T. Nakano. Relational pattern mining based on equivalent classes of properties extracted from samples. PAKDD2008, pp. 582–591, 2008.

[4] J. Motoyama, S. Urazawa, T. Nakano and N. Inuzuka. A mining algorithm using property items extracted from sampled examples, ILP2006, pp.335–350, 2007.