同値なパターンの融合を行う関係型パターンマイニング手法
犬塚研究室　知能系
No. 19115146　美濃羽俊介

1　はじめに

データマイニングとは, 大量のデータから自明でなく有意味な情報を抽出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, その手法の一つ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．本研究では従来の Mapix を発展させて分子構造が同じ原子パターンを繰り返すような同じパターンを繰り返すパターンを抽出できるようにする．

2　Mapix の既存手法

Mapix が組み合わせるアイテムは二種類存在しそれぞれ**性質アイテム**, **分子アイテム**という. 家族関係のデータベースにおいて事例 g(a) 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．

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

これは “a が c という孫息子を持つ” という事実であり, これは有意味であると考えられる. 述語の中で特徴を表す male のような述語をただひとつ持ち, 事例から他の述語を用いてその述語にたどり着くことのできる述語の組を**性質**と考える．性質の定数を変数に置き換えたものを**性質アイテム**と呼ぶ. 性質アイテムの組み合わせの内頻出なもののもとになった性質を組み合わせ変数化したものを**分子アイテム**と呼ぶ, 例えば, 以下のような性質 $p_1,\ p_2$ がある．

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

これらを変数化し性質アイテム $i_1,\ i_2$ を得る．

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

i_1 は “息子をもつ”，i_2 は “孫息子を持つ” という意味である．他方分子アイテムでは $p_1,\ p_2$ をまず組み合わせた後にこれを変数化し次の分子アイテムを得る．

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

これは “これは息子がいてその息子に息子がいる” という意味になる. 性質アイテムや分子アイテムはさらに論理積で組み合わせることができ Apriori 法でマイニングされる. 次に簡単に Mapix の手順を示す

1. 事例から性質を全て抽出する
2. 性質を変数化しパターンとなる**性質アイテム**にする
3. 性質アイテムの頻出な組み合わせを枚挙する
4. 頻出な性質アイテムから**分子アイテム**を生成する
5. 性質アイテムと分子アイテムの頻出な組み合わせを枚挙する

3　提案手法

Mapix **の問題点**Mapix では同値なアイテムの組み合わせを枝刈りすることで高速化を行なっているため同値なアイテム同士の分子アイテムが作られなかった. しかしそれらの中にも有用なものが存在する. 例えば b と c が兄弟である在るという述語 s(b, c) を定義する. これを判定述語とする性質

$$p_2 = \mathsf{g(a)} \leftarrow \mathsf{p(a,b)} \wedge \mathsf{p(a,c)} \wedge \mathsf{s(b,c)}$$
$$p_3 = \mathsf{g(a)} \leftarrow \mathsf{p(a,c)} \wedge \mathsf{p(a,d)} \wedge \mathsf{s(c,d)}$$

を変数化したアイテムは同値であるがこの 2 つを組み合わせて変数化した

$$i_2^2 = \mathsf{g}(A) \leftarrow \mathsf{p}(A,B) \wedge \mathsf{p}(A,C) \wedge \mathsf{p}(A,C) \wedge \mathsf{s}(B,C) \wedge \mathsf{s}(C,D).$$

は “子供が三人いてそれぞれが兄弟である” という意味を持つ. このような同値な性質アイテム同士を融合したアイテムを**同素アイテム**と呼ぶことにする. 同素アイテムを分子アイテムを生成する時と同じ時に生成しようとすると同値なアイテムの枝刈りを行うことが出来ず効率が悪い. そのため, 頻出な性質アイテムを組み合わせる前に**同素アイテム**を抽出する. 性質アイテムの融合を繰り返し最低サポート以上の同素アイテムをすべて得る. 次に提案手法の手順を示す

1. 性質アイテムを得る
2. 性質アイテムから同素アイテムを作る
3. 性質アイテムと同素アイテムの頻出な組み合わせを枚挙する
4. 分子アイテムを得る
5. アイテムの頻出な組み合わせを枚挙する

4　実験結果

実験には人工的な家族関係のデータを最小支持度 40 %で用いた (表 1)．これにより [1] の手法より多くのパターンを抽出できることが確認された.

表 1: 家族関係を用いた実験のパターン数

使用事例数	5	8	10
[1] による手法	17526	18148	19495
提案手法	23621	29280	29299

5　まとめ

本研究では, 関係データマイニングの手法である Mapix に注目し同値なアイテムの融合を用いて同素アイテムを生成し同じパターンを重ねてゆくようなパターンを構成するアルゴリズムを開発した.
しかしながら, 当初の目的であった分子構造のような複雑なデータに対して適応すると組み合わせ爆発が起こり現実的な時間で終わらないという課題が在る. 今後の課題として, より効率的に分子アイテム, 同素アイテムの融合を行う方法を考える方法が必要となる.

参考文献

[1] Y.Nakano, N.Inuzuka : “ Multi-Relational Pattern Mining Basend-on Combination of Properties with Preserving Their Structure in Examples ”, ILP2010, LNAI vol.6489, Springer, pp.181-189, 2010.