3U-7

関係型データマイニングにおける包摂関係を考慮した極小生成系選択

西尾 典晃† 武藤 敦子† 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

形式概念分析は，対象集合とそれが持つ属性集合との間の関係を分析する。関係を持つ対象集合と属性集合が互いに極大となる対を形式概念という。形式概念はパターンマイニングの分野での飽和パターンと数学的に同一であり，この枠組みで冗長性のないパターンの枚挙等の研究が行われている。

また一方で極小生成系とは，形式概念を特徴付ける極小な属性の組である。Dong らは極小生成系から冗長なものを除いた SSMG を提案した。本論文では，複数の関係表に跨がる知識を発見する関係型データマイニング (MRDM) における冗長でない極小生成系を提案する。

本論文では MRDM の Mapix という手法が扱うパターンに着目し，MRDM を形式概念分析の問題に帰着する。さらにアイテム間の包摂関係に基づき，極小生成系について冗長でない簡潔なパターンを定義する。

2 形式概念分析

対象の集合 G と属性の集合 M と二項関係 $I \subseteq G \times M$ が与えられたとき，三つ組 $\mathbb{K} := (G, M, I)$ を文脈といい，形式概念分析 [2] ではこれを分析の対象とする。

二つの集合 G と M に対して任意の二項関係 $I \subseteq G \times M$ を定め，次の写像 I を定義し，形式概念を定める。I はある集合から，その集合のすべての元と関係を持つような元の集合へ移す。

$$X \mapsto X^I := \{m \in M \mid (g,m) \in I \text{ for all } g \in A\} \text{ for } A \in G,$$
$$Y \mapsto Y^I := \{g \in G \mid (g,m) \in I \text{ for all } m \in B\} \text{ for } B \in M.$$

定義 1 (形式概念) 二つ組 (X, Y) は $X \subseteq G, Y \subseteq M, X^I = Y, X = Y^I$ を満たすとき文脈 $\mathbb{K}$ の形式概念という。 □

定義 2 (概念束) 順序関係 $(X_1, Y_1) \leq (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2$ $(\Leftrightarrow Y_2 \subseteq Y_1)$ に対して，形式概念は完備束を作る。これを $\mathbb{K}$ における概念束 $\underline{\mathfrak{B}}(\mathbb{K})$ という。
$\underline{\mathfrak{B}}(\mathbb{K}) := \{(A, B) \mid A \subseteq G, B \subseteq M, A^I = B, A = B^I\}$ □

形式概念 (X, Y) の対象集合 X を外延，属性集合 Y を内包という。また “形式概念” はその概念という名前から次のような解釈ができる。つまり集合 G の要素は “対象”，集合 M の要素は “属性” であり，$(g, m) \in I$ は「対象 g は属性 m を持つ」と読める。概念束はそのような概念的な構造を良く反映している。

On Computing Minimal Generators in Multi-Relational Data Mining with respect to θ-subsumption
†Noriaki NISHIO †Atsuko MUTOH †Nobuhiro INUZUKA
†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

例 1 表 1 は形式文脈を示している。例えば $t_1t_2t_3t_5^I = cdg$ かつ $cdg^I = t_1t_2t_3t_5$ から，二つ組 $(t_1t_2t_3t_5, cdg)$ は形式概念である (本論文では集合の中括弧は省略する)。図 1 の概念束は文脈におけるすべての形式概念を半順序に従ってハッセ図として表したものである。(グラフの頂点は内包を示す)

Tid	a	b	c	d	e	g	h	i
t_1	×	×	×	×	×	×	×	×
t_2	×		×	×		×		
t_3		×	×	×		×	×	×
t_4	×	×		×			×	×
t_5		×	×		×	×	×	×

表 1: 形式文脈

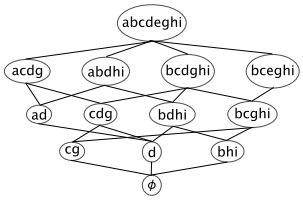

図 1: 表 1 における概念束

3 MRDM と形式概念分析の関係

MRDM で扱うパターンは下記のように一階述語論理で表現される。1 つの述語は 1 つの関係表に対応している。式 (*) の後件に関する関係表において，前件のような複数の関係表に跨るパターンを満たす事例のうち頻出なものを枚挙する。

$$p(X) = head(X) \leftarrow a(X,Y), b(Y,Z), c(Z). \quad (*)$$

MRDM は形式概念分析とは独立した分野であるが，形式概念分析で数学的に定義した対象と属性を各々MRDM における事例集合とパターンの集合に帰着することができる。

MRDM は対象 (上式 (*) では関係 p を満たす事例) と関係するパターンを枚挙している。 このとき，パターンの拡張はリテラルを追加することにより行われる。これを直接，形式概念分析の属性として表現することは難しい。

Mapix[3, 4] は「孫娘を持つ」というような意味としてまとまりのある論理式の集合を “基本パターン” という 1 つの単位としてマイニングを行う手法である。さらに Mapix はボトムアップに探索を行うため，事例が満たすものに限ってパターンを生成する。この 2 つの特徴は元々MRDM のパターンの表現空間が厖大であった問題に対して基本パターンとその組合せに限定するという解を与えた。またこのことは形式概念分析と親和性が高い。つまり，“意味のあるまとまり” を事例が満たす数に限定して枚挙を行っているためパターンが予め定まっており，文脈の形として帰着させることができる。

このような事由から，MRDM の中でも特に Mapix におけるパターンの表現空間を元にして形式概念分析の対象とする。

4 極小生成系と SSMG

本論文では，形式概念を特定する重要な属性の組合せである極小な属性集合に着目する。

定義 3 (極小生成系) 形式文脈 $\mathbb{K}$ 上の形式概念 (X, Y) において，$P \subseteq M$ が極小生成系であるとは，$P^I = X$ でありかつ $\forall P' \subset P$ において $P' \neq X$ となる場合である。 □

例 2 表 1 のような形式文脈が与えられたとき，形式概念 $X = (t_1t_3t_5, bcghi)$ の内包 $bcghi$ に対する極小生成系は bc, bg, ch, ci, gh, gi の 6 つである。この属性を持つような対象集合は X の内包のいずれの属性も含んでいる。

一般にある形式概念に対して複数の極小生成系が存在する。その中で，冗長なものを除き代表的な極小生成アイテム集合のみを選ぼうとしたものが Dong らが提唱した Succinct System of Minimal Generators(SSMG)[1] である。

概念的同値という同値関係を新しく定義することで形式概念に対する極小生成系を同値クラスに分け，クラス毎に SSMG が定められる。

定義 4 (概念的同値) $X, Y \subseteq M$ が以下のいずれかを満たすとき形式概念 C において概念的同値であるといい，$X \approx_C Y$ とかく。

1. X, Y ともに，C より一般な C' の極小生成系である。
2. X, Y は，$Z \approx_C Z'$ と $W \subseteq M$ に対し，
   $X = Z \cup W \approx_C Z' \cup W = Y$ となる。 □

定義 5 (SSMG) 形式文脈の形式概念 $C = (A, B)$ に対する *SSMG* を次のように定める。

1. C が $C \leq (G, \emptyset)$ となる極大の形式概念であるとき，C の極小生成系の概念的同値による同値類から，辞書順で最小の元を *SSMG* とする。
2. 1. 以外では，C に関する極小生成系の概念的同値による同値類から，C より大きい形式概念の *SSMG* のいくつかの和集合となる極小生成系が *SSMG* である。□

5 MRDM における極小生成系

MRDM におけるパターン p は式 (*) のような節である。後件部の述語 $head$ を満たす事例を対象，対象が持つ前件部を属性として形式概念分析の枠組みを適応する。MRDM でのパターンはそのパターン間に包摂関係という依存関係が存在する。本論文ではこの包摂半順序関係に基づいた，冗長性のない極小生成系 Logical Minimal Generators(LMG) を提案する。

定義 6 (包摂) 節 C, D が $C\theta \subseteq D$ を満たす代入 θ を持つとき，節 C は D を包摂するといい，$C \succeq D$ と書く。 □

定義 7 (LMG) *LMG* は $\mathbb{K}$ における任意の形式概念 C に対して次の条件を満たす要素からなる。

1. $C \leq (G, \emptyset)$ が極大な形式概念であるとき，*LMG* は C 上の概念的同値による同値類の中から包摂順序で極小なものである。
2. 1. 以外では，各々の形式概念 C の概念的同値による同値類において，*LMG* は $C' \geq C$ の *LMG* のいくつかの和集合となる極小生成系である。 □

飽和パターンに対する極小生成系に含まれるような属性は，その属性を持つ対象集合の大きさがが他の属性のそれよりも小さい。この事実は極小生成系に含まれるような属性が他の属性よりも特殊な属性であることを示している。この観測から，LMG は極小生成系の中の論理的な依存関係においてより特殊なものを残している。図 2 は集合枚挙木を走査することで LMG を枚挙するアルゴリズムである。

```
LMG_Miner(TDB, sup_min):
  input   TDB     : トランザクションデータベース;
          sup_min : 最低支持度;
  output  LMG     : 冗長でない極小生成系;

 1.  let LMG := ∅;
 2.  let LC := { すべてのトランザクションに現れるアイテム集合 };
 3.  call DFS(H := ∅, T := I − LC, LC);
 4.  return SSMG;
 5.
 6.  DFS(H, T, LC)
 7.  if sup(H) < sup_min return;
 8.  for each x ∈ T
 9.      if sup(H ∪ {x}) = sup(H)
10.          let T := T − {x}, LC := LC ∪ {x};
11.  if (H : LC, sup(H)) が新しい形式概念
12.      add (H : LC, sup(H)) を SSMG に追加;
13.  else 散乱物を削除;
14.  for each p ∈ LC;
15.      H ⪰ p ならば p を LC から削除;
16.  H を LC の mingen として追加;
17.  for each x ∈ T
18.      let H_x := H ∪ {x} and T_x := {y ∈ T | y > x};
19.      call DFS(H_x, T_x, LC);
```

図 2: LMG_Miner：LMG 枚挙アルゴリズム

6 実験及び考察

Mapix の出力したパターンと SSMG，LMG とのパターン数の比較実験を行った。実験では突然変異性を持つ化学物質に関するデータベース Mutagenesis-Bonds を使用した。表 2 は区間 [0.1,0.9] の最低支持度におけるそれぞれの手法が出力したパターン数のグラフである。LMG は他の手法よりもパターン数を抑えられていることが分かる。LMG のパターン数は極小生成系を圧縮できているといえる。ただし，LMG は包摂関係に基づいてパターンの選択を行っているため，元の極小生成系の情報をすべて保っていないことに注意する。

$\sup_{\min}$(%)	90	80	70	60	50	40	30	20	10
Mapix	336	360	614	721	721	925	1467	2948	6630
SSMG	9	9	13	16	16	19	31	67	149
LMG	6	6	9	12	12	14	25	58	137

表 2: 出力されたパターン数の比較

参考文献

[1] Dong, G., et al.: LNCS, 3453 , pp.175-187, 2005.

[2] Ganter, B. and Wille, R.: Formal Concept Analysis: Mathematical Foundations, Springer, 1999.

[3] Motoyama, J., et al.: LNCS, 4455, pp.335-350, 2006.

[4] Nakano, Y. and Inuzuka, N.: LNCS, 6489, pp.181-189, 2010.