計算理論とビッグデータ
ワークショップ

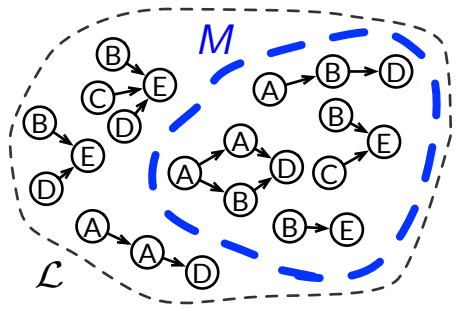
Pattern Structure Analysis for Episode Mining

ポスター内容のソース
発表済:CLA2014
投稿中: IDA Journal
投稿中: ICFCA2015

大滝 啓介(京大/DC2: ootaki@iip.ist.i.kyoto-u.ac.jp), 山本 章博 (京大)

研究背景・目的

- 頻出エピソード発見 [Manilla+,97][Katoh+,09][Tatti+,12]など
  - イベント列 $\mathcal{S}$ からの有向グラフ列挙 (系列データ解析へ応用)
    $\mathcal{S} \in \mathcal{P}(\Sigma)^+ : \{A, B\} \to \{B, D, E\} \to \{B, C, E\} \to \cdots$
- パターン集合発見問題
  - 冗長な全列挙集合の部分集合(~粒子)選択
- 束/情報理論的アプローチによる理論解析/一般化を目指す



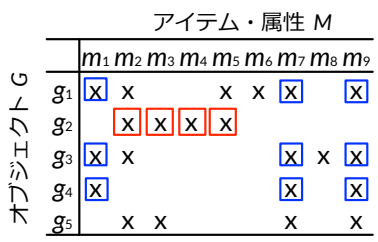
成果

- 菱型/二部エピソードに対する構造概念束の構成手法の提案(基盤としての束構造構築法)
- 情報理論に基づくエピソード選択
- 最小記述長原理(MDL)に基づく束構造抽出の再定式化・既存モデルの一般化を達成
  - 効率的なアルゴリズム等は研究途中

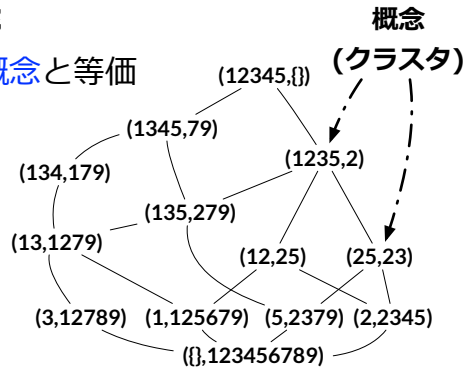
形式概念解析 / アイテム集合束

(DM分野では [Uno+,03] [Pasquier+, 00], FCA等)

- 極大二部クリークを用いた閉包性(closedness)の定義
  - 基本的であり高速に列挙可能
- 形式概念解析(FCA)における概念と等価



(g_{134},m_{179}) はm_1, m_{17}等を含む
(g_2,m_{2345}) は m_{23}等を含む



オブジェクト→アイテム集合
$\delta : G \to \mathcal{P}(M)$

- 計算方法

$$A' = \{m \in M \mid \forall g \in A, (g, m) \in I\} = \bigcap_{g \in A} \{m \mid (g, m) \in I\}$$

$$B' = \{g \in G \mid \forall m \in B, (g, m) \in I\} = \{g \mid \delta(g) \subseteq B\}$$

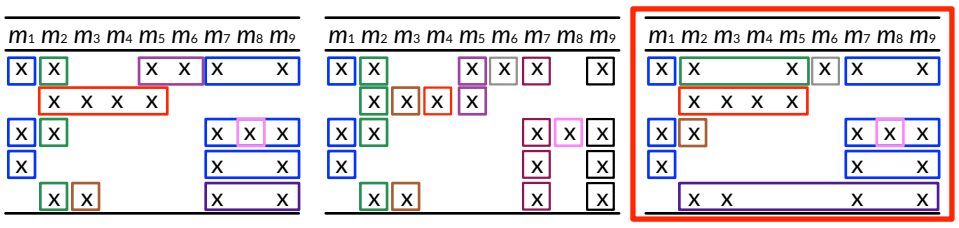
情報量基準と部分集合選択

既存研究 [Vreeken+,06,11] など

- 閉包性は実用的に不十分な場合がある (~ 解が多過ぎる)
- 部分集合の選択 (≒ 部分束の選択)
  - 概念 (~タイル) と符号 (~色) によるデータの符号化
  - 利用するタイル・データ符号化の複雑さを同時に評価
- 二段階最小記述量原理 (Two-step/Crude MDL principle)

$$\arg\min_{\text{ColorTile}} L(I, \text{ColorTile}) = L(I|\text{ColorTile}) + L(\text{ColorTile})$$

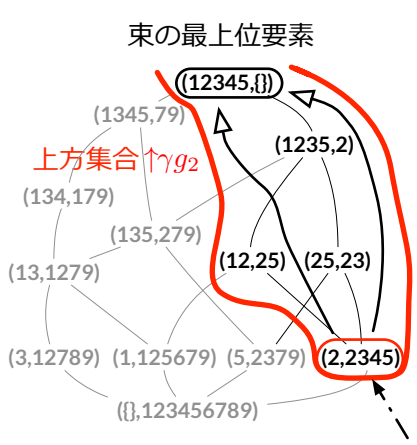
Best



- Krimpアルゴリズム: 逐次的に色付タイルを追加・テスト
- Slimアルゴリズム: MDLを評価しながら探索・列挙

束構造を利用した定式化

- 部分束選択問題を二段階最小化記述量原理を用いて解く
  - 束構造を構築できる全てのデータに対して応用可能
- 問題点: 全ての色付きタイルは束上に概念として出現しない
  - Krimpアルゴリズムの定式化を束の上で最定式化

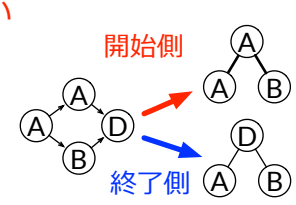


オブジェクト g_2=m_{2345} (FCAではオブジェクトコンセプト γg_2)

- アイデア
  - 接点(概念)間の上下関係を利用
  - パス $\uparrow\gamma g_2 \to (g_{12345}, \{\})$ に着目
  - 複数回通る点でメモ化できそう
- 拡張モデル: 色タイル + 編集操作
  - 編集操作: タイル拡張/色書換
  - → パスをたどるための操作
  - → 束を用いた等価なモデルを達成

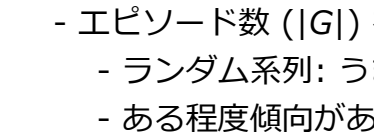
構造概念解析 / 部分構造(集合)束

- 構造データは閉包性はパターンの形に依存して様々で自明ではない
- Pattern structure [Ganter+01]
  - FCAの属性 M を拡張する，しかし全列挙したくない
  - 構造データの部分特徴による閉包性の定義
- アイデア (1) オブジェクト G から特徴抽出
  $\delta : G \to D$　例)有向グラフ ⟼ 高さ1の木(星)の列

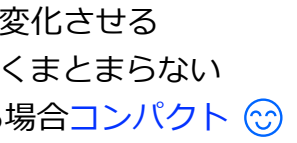


- アイデア (2) 部分特徴の比較/埋め込みの定義

可能な限りラベルを残す



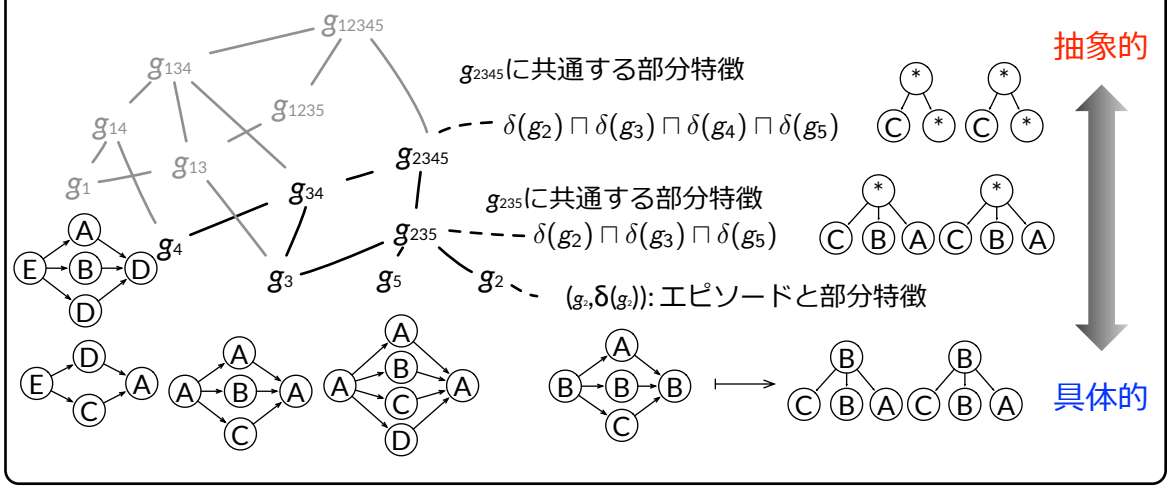
一般化星埋め込み(*や階層/クラスタなどを考慮可能)



$s_1 \sqsubseteq s_2 \iff s_1 \sqcap s_2 = s_1$
(e.g. $s_1 \subseteq s_2 \iff s_1 \cap s_2 = s_1$)

- アイデア (3) FCAに倣った閉包性の定義・構造概念束の構築

$A^\dagger = \sqcap_{g \in A} \delta(g)$　共通する部分特徴の計算
$d^\dagger = \{g \in G \mid d \subseteq \delta(g)\}$　与えられた特徴を持つオブジェクトの集合を探索



実験 I

- 構造概念束の概念 (~クラスタ) を観察
  - ランダム列 (Synth) /MLB投球ログ (M1,M2) から菱型 (D) /二部エピソード (B) を列挙
  - エピソードから星の列を抽出して束を構築
    - エピソード数 ($|G|$) を変化させる
      - ランダム系列: うまくまとまらない
      - ある程度傾向がある場合コンパクト 😊
- 星の列による束構築は列挙よりかなり遅い 😔

データ

Name	$\lvert\mathcal{S}\rvert$	$\lvert\Sigma\rvert$	$\lvert\mathcal{L}_1\rvert$	$\lvert\mathcal{L}_2\rvert$
M1	268	6	832	1093
M2	971	6	971	1397
Synth	392	20	392	1139

概念数の変化

Name	Top 300	ALL	Ratio
M1-D	316	370	1.171
M2-D	372	461	1.239
Synth-D	225	252	1.120
M1-B	1663	2024	1.217
M2-B	1687	2588	1.534
Synth-B	531	1668	3.142

実験 II

※**符号化:** ある星 s を別の星 t から編集操作で生成するために必要な操作に符号を与える(e.g. 色)

- 構造概念束を用いたMDL計算
  - 編集操作: ラベル * の書換, 子節点挿入
  - 符号長は確率に応じた長さ
    $L(c) = -\log_2(\mathbf{Pr}(c))$
  - 星は中心と周辺で別々に符号化して結合

*をAに書換 & 子Bを挿入する

$\mathbf{Pr}(c)$は離散事象 c のデータ中の経験確率

- 実験 I のデータの一部を利用
  1. Greedyに構造概念を選択 😊
  2. MDLの値を評価して更新
  3. 最後に獲得した概念個数を観察

	初期	最後に残った個数	
|Concepts |	|Init |	|Last |	Ratio
218	77	26	0.302
309	140	30	0.234
336	153	6	0.045

- (小規模データでは) ごく一部分だけの概念を選択することに成功 😊

主な参考文献

Buzmakov+, The representation of sequential patterns and their projections within FCA (LML2013), **Ganter+**, Pattern structures and their projections (ICCS2001), **Pasquier+**, Discovering frequent closed itemsets for association rules (ICDT1999), **Uno+**, LCM: An Efficient algorithm for enumerating frequent closed itemsets(FIMI'03), **Vreeken+**, Krimp: mining itemsets that compress (DMKD 23(1), 2011)

現在の課題

- 得られる部分集合に関して ラベル付き/検定 による検証
- 得られる部分集合の性質 / 効率的アルゴリズム の研究