A-028

相関ルール生成を目的とする極小生成子列挙アルゴリズムの空間計算量の改善

望月翔悟 *　　岩沼宏治 †

1　はじめに

相関ルールはデータ中の事象（アイテム集合）の共起・反共起の関係を記述したものである．例えば，正ルール $X \Rightarrow Y$ は，X が生起するとき Y も共起することが多い現象を表しており，右負ルール $X \Rightarrow \neg Y$ は X が起きると Y は共起しないことが多いことを表すものである．データから抽出できる相関ルール，特に負ルールの数は極めて多く，何らかの圧縮処理が必要である．[2] では極小生成子を用いたルール集合の可逆圧縮法を提案し， [3] では極小生成子を用いた正負の相関ルールの直接列挙法を考察している．このたえ，極小生成子の高速列挙は正負の相関ルールマイニング計算において重要である．

先行研究 [1] では，まず初めにデータベースから飽和アイテム集合を抽出し，その飽和集合から極小生成子列挙する手法について考察している．列挙の高速化のために，検査済み済みアイテム集合の全てをハッシュ表に登録して，支持度計算を行わずに極小生成子を列挙する手法を提案している．列挙計算の大幅な高速化は可能になったが，ハッシュ表のサイズが非常に大きくなってしまう問題があった． そこで本研究では，ハッシュ表を極力用いずに飽和アイテム集合から極小生成子を列挙する手法を提案する．

表 1　データベース

TID	トランザクション
1	ABCD
2	ABCD
3	ABC
4	A
5	B
6	D

表 2　飽和集合

飽和 ID	飽和集合	極小生成子
1	A:4	A
2	B:4	B
3	D:3	D
4	ABC:3	AB, C
5	ABCD:2	AD, BD, CD

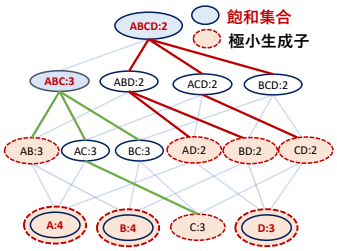

図 1　極小生成子の探索空間

2　準備

$\mathcal{I} = \{A_1, A_2, \ldots, A_n\}$ をアイテムの全体集合とする．**トランザクション** t とは $t \subseteq \mathcal{I}$ なるアイテム集合のことである．**データベース** $\mathcal{D}$ をトランザクションの多重集合と定めるとき，$\mathcal{D}$ 中のアイテム集合 X の**支持度** $\sup(X)$ を $\sup(X) = |\{t \in \mathcal{D} \mid X \subset t\}|$ と定める．また**最小支持度** ms $(0 < ms \leq 1)$ が与えられたとき，$\sup(X) \geq ms$ なる X を**頻出**アイテム集合と呼ぶ．以下では支持度 k のアイテム集合 $\{A_1, \ldots, A_m\}$ を単に $A_1 \ldots A_m : k$ と略記する．

定義 1　アイテム集合 X が**飽和**しているとは，条件 $X \subsetneq X'$ かつ $\sup(X) = \sup(X')$ を満たす X' が存在しない場合しない場合を言う．X の**生成子**を，条件 $Y \subseteq X$ かつ $\sup(Y) = \sup(X)$ を満たすアイテム集合 Y と定める．X の生成子 Y が**極小**であるとは，$Y' \subsetneq Y$ を満たす X の生成子 Y' が存在しない場合を言う．

アイテム集合の（極小）生成子は一般に複数存在することに注意して頂きたい．先行研究 [1] では，データベース $\mathcal{D}$ から極小生成子を列挙するために，まず D から頻出な飽和集合を全て抽出する．次に，各飽和集合 X の部分集合 Y が生成子か否かを判定し，生成子であった場合には極小性の判定を行う． Y が X の極小生成子であれば，Y と X の対を記録する．例えば，表 1 のデータベースと最小支持度 $ms = 1/3$ に対して，飽和集合と極小生成子の対は表 2 のようになる．また図 1 は，表 2 の極小生成子を列挙する探索空間の一例である．飽和集合を根とする複数のトップダウン型の探索木から構成されている．

定義に基づいて生成子/極小性を判定するには，支持度の計算が必要になるが，その計算は非常に重い．そのため先行研究 [1] では，以下の定理 1 に基づき，支持度が不用な判定計算法を開発している．またその実現のために，探索したアイテム集合の全てをハッシュ表に記録している．極小性の判定もこのハッシュ表を用いれば，高速に計算できる．

定理 1（生成子判定原理 [1]）　飽和集合 X とその任意の部分集合 Y に関して以下が同値である

1. $\sup(X) = \sup(Y)$，即ち Y は X の生成子である
2. $Y \subseteq Y' \subsetneq X$ を満たす飽和集合 Y' が存在しない

3　列挙アルゴリズムの空間計算量の改善法

先行研究 [1] で構築するハッシュ表の最大空間計算量は $O(2^{|I|})$ であり，非常に大きい．本論文では，この空間計算量を改善するために，探索したアイテム集合の全てを登録したハッシュ表を構築せずに，生成子と極小性の判定を行う手法を提案する．

* 山梨大学大学院医農工学総合教育部コンピュータ理工学コース
† 山梨大学大学院総合研究部

表 3　バケット型垂直配置表

アイテム	$n=1$	$n=2$	$n=3$	$n=4$
A	1	null	4	5
B	2	null	4	5
C	null	null	4	5
D	3	null	null	5

3.1　生成子判定

定理 1 の 2 は，以下の 2’ に言い換えることができる．

2’. $|Y| \leq |Y'| < |X|$ なる任意の飽和集合 Y' に対して，Y は Y' の部分集合とはならない．

本論文では，2’ の判定をバケット型垂直配置表を用いて効果的に行う．バケット型垂直配置表とは，各アイテムが含まれる飽和集合を，飽和集合の大きさごとに記録（バケット）した配置表である．例として，表 2 のバケット型垂直配置表を表 3 に示す．表 3 中の n は飽和集合の大きさを表している．飽和集合 ABC（飽和 ID=4）はアイテム A，B，C を含むので，配置表の A，B，C の $n=3$ のバケットに ID=4 が記録されている．

バケット型垂直配置表を用いて，飽和集合 X のある部分集合 Y が X の生成子であるか判定を行う．$|Y| \leq n < |X|$ なるバケットを参照して，Y を部分集合として含む飽和集合が存在するか否か判定することにより，生成子判定を行う．例えば，飽和集合 ABCD の部分集合 BC は，表 3 では，$2 \leq n < 4$ を満たす n のバケットを参照すると，BC は飽和 ID=4（ABC）の部分集合であることが分かるので，BC は ABCD の生成子でないと判定できる．

3.2　極小生成子判定

本文では，生成子の極小性を判定するために以下の定理を導いた．

定理 2（極小性判定原理） 飽和集合 X とその生成子 Y に関して以下が同値である．

1. Y は X の極小生成子である
2. Y の任意の真部分集合 Z は，$N \subsetneq X$ を満たすある飽和集合 N の極小生成子である

紙面の都合上，証明は割愛する．上の条件 2 は，小さい飽和集合から順に極小生成子を探索列挙し，見つかった極小生成子だけをハッシュ表に登録することによって，高速に判定できる．例えば，表 2 の状況において，飽和集合 ABC の生成子 AB について考える．AB の真部分集合 A，B はそれぞれ極小生成子で，かつ対の飽和集合が ABC ではないため，AB は ABC の極小生成子であると判定できる．

4　実験と考察

先行研究 [1] との比較実験の結果を示す．紙面の都合上，データセット accidents に対する使用メモリ量と実行時間のグラフだけを図 2 と 3 に示す．図中の「支持度なし」は探索を記録するハッシュ表を用いて支持度計算を行わない場合 [1] である．「支持度あり」は垂直配置表を用いて持度比較を行った場合の結果である．

図 2 より，提案手法は先行研究よりも少ないメモリで極小生成子列挙を行えいることが分かる．実行時間は図 3 より，最小支持度が大きい場合は提案法は先行研究よりも高速であるが，最小支持度が小さい場合には低速であることが分かる．最小支持度が小さい場合，複数の飽和集合の極小生成子を列挙を行う過程で，同一のアイテム集合を何度も判定する必要が生じる．提案手法ではその度にバケット型垂直配置表を用いた生成子判定を行っており，冗長な重複計算が発生しているためと考えられる．

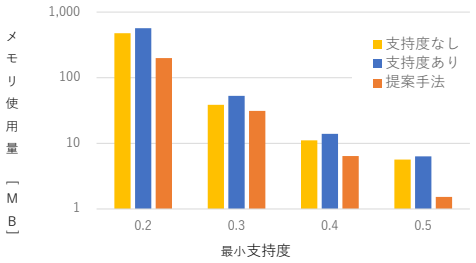

図 2　accidents 使用メモリ

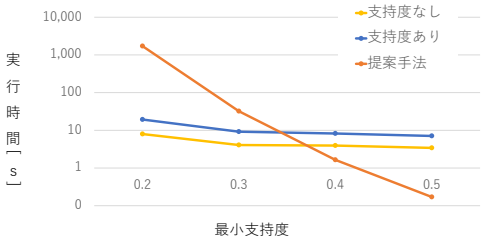

図 3　accidents 実行時間

5　まとめと今後の課題

本稿では全ての探索済みアイテム集合を記録せずに，バケット型垂直配置表と作成中の極小生成子を記録したハッシュ表を用いて極小生成子を列挙する省メモリ型アルゴリズムを提案した．今後は高速化するために，重複再計算を防止する手法の組込みについて研究していく予定である．

謝辞： 本研究の一部は JSPS 科学研究費補助金 (No.19K12096, No.22K12165) の援助を受けている．

参考文献

[1] Koji Iwanuma, Kento Yajima and Yoshitaka Yamamoto: Enumerating Minimal Generators from Closed Itemsets: Toward Effective Compression of Negative Association Rules. Proc. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp.1-7 (2021)

[2] 岩沼宏治，佐生隼一，黒岩健歩，山本泰生: 負の相関ルール集合の極小生成子に基づく圧縮表現，情報処理学会論文誌，57 巻 8 号，pp. 1845-1849，(2016)

[3] 谷島健斗，岩沼宏治，黒岩健歩，佐生隼一，山本泰生: 極小生成子を用いた負の相関ルール抽出計算の効率化，第 31 回人工知能学会全国大会，4A1-3in1，(2017)