4A1-3in1

# 極小生成子を用いた負ルール抽出計算の効率化

Efficient Mining for Negative Association Rules Based on Minimal Generators

谷島 健斗 *1 Kento yajima　岩沼 宏治 *2 Koji Iwanuma　黒岩 健歩 *3*4 Yasuho Kuroiwa　佐生 隼一 *3 Shunichi Sasho　山本 泰生 *2*5 Yoshitaka Yamamoto

*1山梨大学大学院医工農学総合教育部工学専攻コンピュータ理工学コース
Conputer Science and Engineering Course, Integrated Graduate School of Medicine, Engineering and Agricultural Sciences, University of Yamanashi

*2山梨大学大学院総合研究部
Interdisciplinary Graduate School, University of Yamanashi

*3山梨大学大学院医学工学総合教育部コンピュータ・メディア工学専攻
Computer Seience and Media Engineering, Interdisciplinaly Graduate School of Medieine and Engineering, University of Yamanashi

*5科学技術振興機構 さきがけ
JST PRESTO

In this paper, we propose an efficient method for mining negative association rules, using minimal generators. The number of negative association rules extracted from a transaction database is extremely large, compared with the one of ordinary positive association rules. Therefore, some compression technique for negative rules is an inevitable and rational solution for efficient negative rule mining. A minimal generator is a lossless compressed forms of itemsets. We give an efficient extraction method for the negative rule, which takes the form of a depth-first search over a suffix tree consisting of frequent minimal generators. We also show some preliminary results of evaluation experiments.

## 1. はじめに

本論文では，極小生成子を用いた負の相関ルール抽出手法を提案し，評価実験により本手法の有効性を検証する．相関ルールとは, トランザクションデータベース中に頻繁に共起するアイテム集合の関係を記述したものである．$X$ と $Y$ をアイテム集合とするとき, データベース中で $X$ が出現するトランザクションの多くに $Y$ も出現することを $X \Rightarrow Y$ と表し, 正の相関ルールと呼ぶ．これに対して, 本論文では, $X$ と $Y$ がほとんど同時に出現しない現象を表現する $X \Rightarrow \neg Y$ や $\neg X \Rightarrow Y$ なる形の負の相関ルールを考察する．負の相関ルールは正のルールでは表現が困難な共起関係を記述でき, データから有益な情報を抽出することが可能になる．ただし, 負のルールを抽出するためには, 非頻出なアイテム集合を扱う必要がある．そのため, 正の相関ルールの場合に比べて探索空間が格段に大きく, また抽出されるルールの数も非常に多くなる．井出らは接尾辞木を用いた深さ優先型の負ルール抽出アルゴリズムを提案し, 負ルール抽出の効率化を行った [1]．岩沼らは極小生成子を用いて抽出される負ルールの圧縮を行った [2]．本研究では, 有効な負の相関ルール集合を圧縮して, 効率的に抽出を行うために, 頻出アイテム集合ではなく, 極小生成子を用いて負ルール抽出を効率的に行う計算手法を提案する．はじめに予備的考察として, 簡単な例を用いて提案手法の有効性を示す．さらに, 評価実験の結果, 良い結果を得ることができたので, 報告する．

連絡先: 谷島 健斗，山梨大学大学院工学専攻 (修士課程) コンピュータ理工学コース，山梨県甲府市武田 4-3-11, t13cs053@yamanashi.ac.jp
*4現職:デンソーテクノ株式会社

## 2. 準備

### 2.1 正の相関ルール

$I = \{x_1, x_2, \cdots, x_n\}$ を**アイテム**の全体集合とする時，**トランザクション** $t$ をアイテム集合 $t \subseteq I$ と定める．トランザクションの多重集合を**トランザクションデータベース** $D$ とする．$X$ をアイテム集合とするとき，$X \subseteq t$ となる $D$ 中のトランザクション $t$ を $X$ の出現と呼び，その集合を $D(X)$ と略記する．集合 $X$ の大きさを $|X|$ と表記するするとき，$X$ の $D$ 中の**支持度** $sup(X)$ を $sup(X) = \frac{|D(X)|}{|D|}$ と定義する．**正の相関ルール** (以下，適宜“正ルール”と略記する) を $X \cap Y = \emptyset$ であるアイテム集合 $X$，$Y$ からなる表現 $X \Rightarrow Y$ と定める．$X$ と $Y$ をそれぞれルールの**前件**，**後件**と呼ぶ．正ルールに対する**支持度** sup と**確信度** conf は以下のように定義される．

$$\mathrm{sup}(X \Rightarrow Y) = \mathrm{sup}(X \cup Y) \tag{1}$$

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{sup}(X \cup Y)}{\mathrm{sup}(X)} \tag{2}$$

**最小支持度** $ms$ と**最小確信度** $mc$ とはユーザが支持度と確信度に関して与える閾値である．$sup(X) \geq ms$ を満たす $X$ を**頻出アイテム集合**と呼ぶ．また $sup(X \Rightarrow Y) \geq ms$ と $conf(X \Rightarrow Y) \geq mc$ の両方を満たす $X \Rightarrow Y$ を**有効** (valid) な正の相関ルールと呼ぶ．

### 2.2 負の相関ルール

本論文では, **負の相関ルール** (*negative association rule:* 以下では適宜 “負ルール” と略記する) を考察する．$X$ と $Y$ を $X \cap Y = \emptyset$ とする時, 負ルールとは以下のいずれかの表現である．

- $X \Rightarrow \neg Y$　　(右否定形もしくは後件負形)

- $\neg X \Rightarrow Y$　　(左否定形もしくは 前件負形)
- $\neg X \Rightarrow \neg Y$　　(両否定形)

両否定形 $\neg X \Rightarrow \neg Y$ は, 一般に非常に数が多い. そのため, 両否定形の効率的な抽出は困難である. また, ルールとしての有用性も低いため, 本論文では両否定形を扱わない. 上記の $\neg X$ はアイテム集合の否定表現であり，**負アイテム集合**と呼ぶ. 以下では $C_X$ はアイテム集合 $X$ または負アイテム集合 $\neg X$ のどちらかを表すものとする.

負アイテム集合および負ルールの**支持度** sup と**確信度** conf を以下のように定める.

$$\sup(\neg X) = 1 - \sup(X) \tag{3}$$

$$\sup(X \Rightarrow \neg Y) = \sup(X) - \sup(X \cup Y) \tag{4}$$

$$\sup(\neg X \Rightarrow Y) = \sup(Y) - \sup(X \cup Y) \tag{5}$$

$$\sup(\neg X \Rightarrow \neg Y) = 1 - \sup(X) - \sup(Y) + \sup(X \cup Y) \tag{6}$$

$$\mathrm{conf}(C_X \Rightarrow C_Y) = \frac{\sup(C_X \Rightarrow C_Y)}{\sup(C_X)} \tag{7}$$

以下の 5 つの条件を満たすルール $C_X \Rightarrow C_Y$ を**有効な負ルール**と呼ぶ.

1. $X \cap Y = \emptyset$　　(独立性)
2. $\sup(X) \geq ms$　かつ　$\sup(Y) \geq ms$ (前件と後件の頻出性)
3. $\sup(X \Rightarrow Y) < ms$ (無矛盾性)
4. $\sup(C_X \Rightarrow C_Y) \geq ms$ (ルールの頻出性)
5. $\mathrm{conf}(C_X \Rightarrow C_Y) \geq mc$ (ルールの確信度)

条件 1 は前件と後件の独立性の条件である. 条件 2 は, アイテム集合 $X, Y$ が共に頻出アイテム集合という条件である. 条件 3 は, 無矛盾性の条件であり, 先行研究 [1] で定義したものである. これは, 正ルール $(X \Rightarrow Y)$ が有効であるとき, 同時に負ルール $(C_X \Rightarrow C_Y)$ の抽出を行わない条件である. 条件 4 は支持度条件であり, 条件 5 は確信度条件である. 本論文では, 全ての有効な負ルールを抽出する計算問題に取り組む.

## 3. 先行研究

本章では, 本研究の先行研究にあたる負ルール抽出法および負ルール集合の圧縮表現に関する研究について述べる.

### 3.1 飽和アイテム集合と極小生成子

アイテム集合 $X$ に対して, $X \subset X'$, $X \neq X'$ かつ $\sup(X)$=$\sup(X')$ を満たす $X'$ が存在しないならば, $X$ を**飽和アイテム集合** (*closed itemset*) と呼ぶ.

また $X$ が $X'$ に対して，$X \subseteq X'$ かつ $\sup(X') = \sup(X)$ を満たすならば, $X$ を $X'$ の**生成子** (*generators*) と呼ぶ. 生成子は一般には複数存在するので, その中でより小さな生成子が存在しないものを**極小生成子** (*minimal generators*) [3] と呼ぶ.

### 3.2 頻出アイテム集合と接尾辞木を用いた負ルール抽出

先行研究 [1] のルール抽出アルゴリズムは, 頻出アイテム集合から接尾辞木を作成し, 左優先深さ優先探索を行い, 負ルール抽出を行う.

ここで, 接尾辞木について説明していく. アイテム間には適当な順序を仮定し、アイテム集合をアイテムの列としてとり扱う. 図 1 は, 接尾辞木 (*suffix tree*) の例を示している. 以下では, アイテム集合 $\{A, B, C\}$ を $ABC$ と略記する. 図 1 ではアルファベット順を仮定している. 各節点 $N_c$ の親は長さが 1 つだけ短い**接尾辞** (*suffix*) を持つ節点 $N_p$ である. 兄弟関係にある節点は順番に左から右へ並ぶ. 接尾辞木上で左優先深さ優先探索を行うと，節点 $N$ を訪問する時点で $N$ の部分集合は全て訪問が完了していることが保障されている [4].

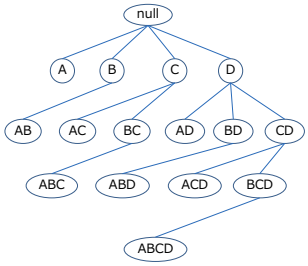


図 1: 接尾辞木

以下に, 先行研究 [1] における, 有効な負ルール抽出アルゴリズムを示す. 擬似コード中のアイテム集合 X, Y は接尾辞木上のある節点のアイテム集合である. *Check_Rule*(X, Y) で $X \Rightarrow \neg Y$ と $\neg Y \Rightarrow X$ が有効なルールであるかチェックしている. また, 同時に適宜枝刈りを行い, 探索の効率化を行っている.

**Algorithm 1** 負の相関ルールの大枠

トランザクションデータベースから頻出アイテム集合（*FIS*) の集合を抽出し, $N$ を抽出したアイテム集合の総和とする;
1: *FIS* の集合から接尾辞木を生成;
2: 接尾辞木上で左優先深さ優先探索で頻出アイテム集合を $FIS_1$, ... ,$FIS_N$ の順に並べる;
3: **for** $i = 1$ to $N$ **do**
4:　　$X = FIS_i$
5:　　**for** $j = 1$ to $N$ **do**
6:　　$Y = FIS_j$
7:　　Check_Rule$(X, Y)$
8:　　**end for**
9: **end for**

ルール抽出規則 Check_Rule$(X, Y)$ での主な工夫点を以下に示す.

- $X \Rightarrow \neg Y$ と $\neg Y \Rightarrow X$ の支持度計算を同時に行う
- 左否定形の確信度の枝刈りに上界関数を用いる
- 右否定形に対して極小なルールを定義し, 接尾辞木の性質を利用して極小でないルールを冗長なルールとして, 生成を禁止している

### 3.3 極小生成子を用いた負ルールの圧縮表現

頻出アイテム集合や正のルールの圧縮には飽和アイテム集合がよく用いられているが, 負ルールの圧縮に用いた場合, 本来抽出すべき負ルールが表現できなくなる現象が生じる. そこで, 先行研究 [2] において, 飽和アイテム集合と対の表現である極小生成子を用いることで妥当な負ルール集合を圧縮できることが示されている. 例を用いて負ルールの圧縮について説明を行っていく.

表 1 にデータベースを示す. ここでは, 右否定形のみを扱っていく. $ms$ を 0.4 としたとき, 表 2 の 6 個のルールが抽出されたとする. 飽和アイテム集合を用いて表 2 のルールを圧縮していく. 飽和アイテム集合は, $AB, B, BCD$ である. 表 2 の

ルールは $AB \Rightarrow \neg BCD$ で表現することができる．しかし，圧縮後のルール $AB \Rightarrow \neg BCD$ は，3.1 で述べた負ルール条件の独立性条件を満たさないため，本来抽出すべき負ルールを表現することができない．そこで，極小生成子を用い 6 個の負ルールを圧縮する．極小生成子は $A, B, C, D$ であるので，表 2 のルールは $A \Rightarrow \neg C$ と $A \Rightarrow \neg D$ の 2 つのルールで表現することができる．この 2 つのルールはどちらも有効な負ルールであり，負ルールの圧縮には極小生成子が有効であるといえる．

表 1: データベース

| TID | アイテム集合 |
|---|---|
| 1 | $AB$ |
| 2 | $AB$ |
| 3 | $BCD$ |
| 4 | $AB$ |
| 5 | $BCD$ |
| 6 | $BCD$ |

表 2: 抽出ルール例

| No | 抽出ルール |
|---|---|
| 1 | $A \Rightarrow \neg C$ |
| 2 | $A \Rightarrow \neg D$ |
| 3 | $A \Rightarrow \neg BC$ |
| 4 | $A \Rightarrow \neg CD$ |
| 5 | $A \Rightarrow \neg BD$ |
| 6 | $A \Rightarrow \neg BCD$ |

## 4. 極小生成子の性質

本研究では，極小生成子の今まで知られていなかった以下の性質を発見した．

**[定理 1]** アイテム集合 $MG$ が極小生成子ならば，$MG$ に含まれている**全て**の部分集合は極小生成子になる．

**証明** $MG_0$ が要素数 $n$ の極小生成子であるとき，その部分集合が全て極小生成子になることを背理法を用いて証明する．$MG_0$ の頻度を $f_0$ とする．また，$MG_0$ より 1 つ要素が少ない集合を $S_1, ..., S_n$ とし，頻度を $f_1, ..., f_n$ とする．このとき $MG_0$ が極小生成子なので，$f_0 > f_1, ..., f_0 > f_n$ となる．ここで，$S_1$ が極小生成子にならないと仮定する．仮定より，$S_1$ には $S_1$ より要素数の少ない部分集合に極小生成子が存在する．その極小生成子を $MG_1$ とすれば，$\alpha \in S_1$ かつ $\alpha \notin MG_1$ となる要素 $\alpha$ が少なくとも 1 つ存在する．$MG_0$ の部分集合 $S_2, ..., S_n$ において，$MG_1$ を含み，$\alpha$ を含まない集合が存在するので，それを一般性を失わずに $S_2$ と仮定する．$S_2 \cup \{\alpha\} = MG_0$ であるため，$S_2 \cup \{\alpha\}$ の頻度は，$f_0$ となる．また一方で，$\alpha \in S_1$ かつ $\alpha \notin MG_1$ で $MG_1$ が $S_1$ の極小生成子であるから，$MG_1$ が出現するトランザクションには $\alpha$ も必ず同時に出現する．これより，$MG_1 \subset S_2$ なので $S_2 \cup \{\alpha\}$ の頻度は $f_2$ に等しい．よって $f_2 = f_0$ となり，$MG_0$ が極小生成子であるという仮定に矛盾．よって，極小生成子に含まれる要素の部分集合は全て極小生成子になる．

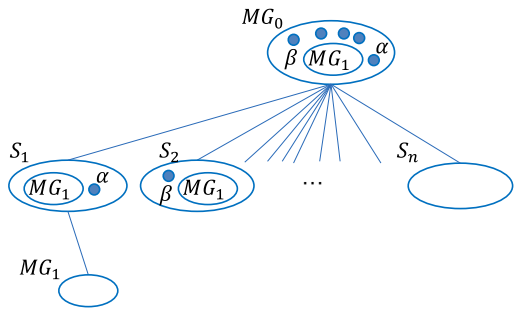


図 2: 証明の概要図

## 5. 極小生成子に基づく負ルール抽出

先行研究 [2] より，極小生成子を用いることで負ルールの圧縮表現ができることが保証できる．また，4 章で述べた性質より，極小生成子だけを用いて接尾辞木を作成することもできるため，先行研究 [1] で提案された手法を用いることができる．そこで本研究では，以下で，提案手法のアルゴリズムは先行研究 [1] のアルゴリズムの頻出アイテム集合の部分を極小生成子に変更した手法の性能について研究を進める．

予備的考察として，表 3 のデータベースから抽出されるルールの数の比較を行った．$ms$ を 0.2, $mc$ を 0.5 と設定した．既存手法で作成された接尾辞木を図 2 に示し，提案手法で作成された接尾辞木を図 3 に示す．図 2 と 3 より，既存手法では大幅に探索空間を削減できる．抽出されたルール数を表 5 に示す．表 5 より提案手法での抽出ルール数は既存手法の約 1/4 倍である．これより実データに対しても，提案手法が有効であることが推測されるので，次に評価実験を行った．

表 3: データベース

| TID | アイテム集合 |
|---|---|
| 1 | $ABCD$ |
| 2 | $ABCD$ |
| 3 | $BCD$ |
| 4 | $DEF$ |
| 5 | $DEF$ |
| 6 | $C$ |
| 7 | $E$ |

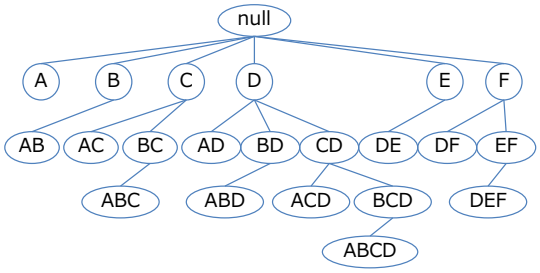


図 3: 頻出アイテム集合を頂点とした接尾辞木

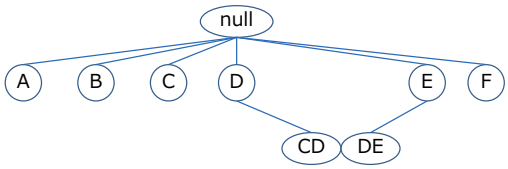


図 4: 極小生成子を頂点とした接尾辞木

表 4: 抽出結果

| | 総ルール数 | 左否定形 | 右否定形 |
|---|---|---|---|
| 既存手法 | 132 | 86 | 46 |
| 提案手法 | 35 | 18 | 17 |

表 7: 提案手法での負ルール抽出の実験結果

| データセット | ms | | sup 検査対 | 総ルール数 | 左否定形 | 右否定形 | 実行時間 (s) |
|---|---|---|---|---|---|---|---|
| retail | 0.002 | 既存 | 2,745,419 | 2,528,923 | 1,804 | 2,527,119 | 9.87 |
| | | 提案 | 2,745,419 | 2,528,923 | 1,804 | 2,527,119 | 9.64 |
| | 0.003 | 既存 | 782,258 | 699,252 | 934 | 698,318 | 4.09 |
| | | 提案 | 782,258 | 699,252 | 934 | 698,318 | 3.99 |
| | 0.004 | 既存 | 289,775 | 255,289 | 604 | 254,685 | 2.16 |
| | | 提案 | 289,775 | 255,289 | 604 | 254,685 | 2.11 |
| mushroom | 0.2 | 既存 | 90,743,656 | 13,346,615 | 12,595,185 | 751,430 | 178.85 |
| | | 提案 | 1,166,448 | 220,262 | 196,276 | 23,986 | 10.53 |
| | 0.3 | 既存 | 1,445,212 | 417,449 | 406,527 | 10,922 | 3.51 |
| | | 提案 | 131,870 | 38,041 | 35,657 | 2,384 | 1.14 |
| | 0.4 | 既存 | 94,134 | 9,800 | 9,372 | 428 | 0.51 |
| | | 提案 | 16,136 | 2,110 | 1,960 | 150 | 0.25 |

## 6. 評価実験

本研究で提案した手法を実装して負の相関ルールを抽出し, 提案手法の効果を測定した. 実験には, Frequent Itemset Mining Dataset Repository [5] から 2 種のデータセットを使用した. 各データセットの詳細を表 6 に示す. #(item) はデータセットに含まれるアイテムの種類数を示し, #(trans) はデータセット中のトランザクションの総数, ave(item) は 1 トランザクション中に出現するアイテムの平均数である. #(FIS) は頻出アイテム集合の総数, #(CIS) は頻出飽和アイテム集合の総数, #(MG) は極小生成子の総数である.

表 5: 実験に使用したデータセットの概要

| データセット | #(item) | #(trans) | ave(item) |
|---|---|---|---|
| retail | 16,470 | 88,162 | 10.3 |
| mushroom | 119 | 8,124 | 23 |

表 6: 抽出数の概要

| データセット | ms | #(FIS) | #(CIS) | #(MG) |
|---|---|---|---|---|
| retail | 0.002 | 2,691 | 2,691 | 2,691 |
| | 0.003 | 1,393 | 1,393 | 1,393 |
| | 0.004 | 831 | 831 | 831 |
| mushroom | 0.2 | 53,663 | 1,197 | 1,607 |
| | 0.3 | 2,735 | 427 | 539 |
| | 0.4 | 565 | 140 | 170 |

最小確信度 $mc$ を 0.4 に固定し, 最小支持度 $ms$ の値を変化させて負ルールを抽出した実験結果を表 8 に示す. 疎なデータセットである retail において, 表 7 より, 頻出アイテム集合, 飽和アイテム集合, 極小生成子の数が同じであることから, 極小生成子によって圧縮が行えないことが示された. 表 8 より, 抽出ルール数は, 提案手法において既存手法と同様な結果が示された. 実行時間についても, 既存手法と同等であることが示された. また, 密なデータセットである mushroom において, 表 7 より, 極小生成子によって大幅に圧縮が行えていることが示された. 表 8 より, 抽出ルール数が大幅に削減された. 実行時間についても, 削減された. 以上の結果より, 提案手法は密なデータセットにおいて, 有効であることが示された. 極小生成子の抽出手法については, 改善の余地があり, 極小生成子の抽出に関する研究の手法 [6] を導入することでさらなる高速化が期待できる.

## 7. まとめ

本研究では, 頻出アイテム集合の圧縮形である極小生成子を用いた効率的な負の相関ルール抽出手法を検討した. ルールを抽出する際のアイテム組合せが減るため, 効率的な負ルールの抽出が可能である. 実際, 評価実験により, 密なデータセットに対して効果を確認することができた. 今後の課題として, 極小生成子の高速抽出と特徴を生かした新たな枝刈り方法の検討する. また, 関連尺度の導入を行う.

## 謝辞

本研究は一部，ISPS 科学研究費補助金（No.16K00298）および JST さきがけの援助を受けている．

## 参考文献

[1] 井出典子, 岩沼宏治, 山本泰生: 負の相関ルールを抽出する高速トップダウン型アルゴリズム, 人工知能学会論文誌, 29 巻 4 号, pp. 406-415, (2014)

[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 1 号, pp. 1-5, (2016)

[3] M. J. Zaki: Mining Non-Redundant As sociation Rules. *Data Mining and Knowledge Discovery*, Vol. 9, pp. 223-248, (2004)

[4] 亀谷由隆, 佐藤泰介: 最小サポート上昇法に基づく上位 $k$ 関連パターンの発見, 人工知能学会データ指向構成マイニングとシミュレーション研究会, SIG-DOCMAS, B101-4, pp.(2-24)-(2-32), (2011)

[5] Frequent Itemset Mining Dataset Repository, <http://fimi.ua.ac.be/>(2017-2-28)

[6] 佐生 隼一, 岩沼 宏治, 山本 泰生, 黒岩 健歩: 負の相関ルールマイニング効率化のための極小生成子の抽出計算, 人工知能学会第 103 回人工知能基本問題研究会, (2017)