正負の相関ルールの妥当性の再考察と正負ルールの高速抽出手法

Reconsideration of Validity of Positive and Negative Association Rules and Their Fast Extraction Methods

雨宮 晶良[1*] 岩沼 宏治[2] 谷島 健斗[1] 山本 泰生[2]
Akira Amemiya[1] Koji Iwanuma[2] Kento Yajima[1] Yoshitaka Yamamoto[2]

[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
[1] Computer 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
[2] Interdisciplinary Graduate School, University of Yamanashi

Abstract: We reconsider validity conditions of positive and negative rules in the association rule mining and also study the adaptability of the statistical correlation measures for positive and negative rules. Based on the validity conditions, we propose an efficient extraction method of top-k association rules using the branch and bound mechanism with several anti-monotonic upper bound functions.

1 はじめに

本論文では、正負の相関ルール [1] の妥当性を再考察し, 新しい定義を提案する. 併せて関連性尺度による負ルールの絞り込みを考察する. それを踏まえた上で, 関連性尺度を用いた分枝限定法による上位 k 個の正負の相関ルール抽出手法を提案し, 実証実験により提案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 $\neg X \Rightarrow Y$, $\neg X \Rightarrow \neg Y$ 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相関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 [1, 12, 13]. ただし, 負の相関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と比べて探索空間が格段に広く, また抽出されるルールの数も非常に多くなる.

井出ら [1] は, 抽出すべき負の相関ルールの性質を考察し, **妥当な負の相関ルール**として定式化した. また, 接尾辞木を用いた深さ優先型と分枝限定法による負ルール抽出法を提案し, 負ルール抽出アルゴリズムの効率化を行った. 黒岩ら [2] は, 関連性尺度として cosine を用いた上位 k 個の負ルール抽出アルゴリズムを提案した. 岩沼ら [4] は, 極小生成子による負ルールの無損失圧縮を提案した. 谷島 [3] は, 極小生成子を用いた負ルールの圧縮と抽出のアルゴリズムを提案した.

本論文では, 井出ら [1, 13] の提案した妥当な負ルールの条件を再考察し, それを踏まえたうえで分枝限定法に用いる幾つかの逆単調な上界関数を新しく考え, 効率的な正負の上位 k 個のルールの同時抽出法を考察する.

*連絡先：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
E-mail:g18tk001@yamanashi.ac.jp

2 準備

2.1 正・負の相関ルール

$I = \{x_1, x_2, \cdots, x_n\}$ 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. **トランザクションデータベース** D をトランザクションの多重集合とする. 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, その多重集合を $D(X)$ と略記する. 多重集合 A の**大きさ**を $|A|$ と表記するとき, $|D(X)|$ を X の D 中の**絶対出現頻度**と呼ぶ. X の D 上の**支持度 (相対出現頻度)** $\mathbf{sup}(X)$ を $\sup(X) = \frac{|D(X)|}{|D|}$ と定義する. **正の相関ルール** (以下, “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. 正の相関ルールに対する**支持度** sup と**確信度** conf は以下のように定義される.

$$\mathrm{sup}(X \Rightarrow Y) = \mathrm{sup}(X \cup Y), \quad \mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{sup}(X \cup Y)}{\mathrm{sup}(X)}$$

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対して与える閾値である. $\mathrm{sup}(X) \geq ms$ を満たす X を**頻出アイテム集合**と呼ぶ. また $\mathrm{sup}(X \Rightarrow Y) \geq ms$ と $\mathrm{conf}(X \Rightarrow Y) \geq mc$ の両方を満たす $X \Rightarrow Y$ を**妥当な** (vaild) 正の相関ルールと呼ぶ.

また本論文では, **負の相関ルール** (以下では “負ルール” と略記) を考察する. X と Y を $X \cap Y = \emptyset$ なるアイテム集合とするとき, 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**)
- $\neg X \Rightarrow Y$ (**左否定形**)
- $\neg X \Rightarrow \neg Y$ (**両否定形**)

上記の $\neg X$ はアイテム集合の否定表現であり, **負のアイテム集合**と呼ぶ. 以下では C_X をアイテム集合 X または負のアイテム集合 $\neg X$ のどちらかを表すものとする. 負のアイテム集合とルールの**支持度 sup** と**確信度 conf** を先行研究 [1, 12, 13] に従い, 以下のように定める.

$$\begin{aligned}
\mathrm{sup}(\neg X) &= 1 - \mathrm{sup}(X) \\
\mathrm{sup}(X \Rightarrow \neg Y) &= \mathrm{sup}(X) - \mathrm{sup}(X \cup Y) \\
\mathrm{sup}(\neg X \Rightarrow Y) &= \mathrm{sup}(Y) - \mathrm{sup}(X \cup Y) \\
\mathrm{sup}(\neg X \Rightarrow \neg Y) &= 1 - \mathrm{sup}(X) - \mathrm{sup}(Y) + \mathrm{sup}(X \cup Y) \\
\mathrm{conf}(C_X \Rightarrow C_Y) &= \frac{\mathrm{sup}(C_X \Rightarrow C_Y)}{\mathrm{sup}(C_X)}
\end{aligned}$$

2.2 飽和アイテム集合と極小生成子

先行研究の負ルール集合の圧縮表現について概説する. あるアイテム集合 X に対して, $X \subset X'$, $X \neq X'$ かつ $\mathrm{sup}(X)$=$\mathrm{sup}(X')$ を満たす X' が存在しない場合, X を**飽和アイテム集合** [5] と呼ぶ. またアイテム集合 X に対して, $X' \subseteq X$ かつ $\mathrm{sup}(X') = \mathrm{sup}(X)$ を満たす X' を, X の生成子と呼ぶ. 生成子は一般には複数存在するが, その中でより小さい生成子が存在しないものを**極小生成子** [5] と呼ぶ.

頻出アイテム集合や正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. しかし, 負ルールの圧縮に飽和アイテム集合を用いた場合, 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 [4]. 先行研究 [4] では, 飽和アイテム集合ではなく極小生成子を用いることで抽出すべき負ルールの集合を圧縮できることを示した.

3 妥当なルール条件の再考察

本節では先行研究で提案されている, 正と負ルールの妥当性条件を考察する.

3.1 先行研究での妥当性の定義

負ルールの妥当性は, 先行研究 [12, 13] では下記の S1〜S4 を満たすことと定義, 提案されていた. 井出ら [1] では更に S5 の無矛盾性を考慮した, 以下の 5 つの条件を妥当なルールの条件として提案した. 支持度, 確信度それぞれの閾値を ms, mc とし, X と Y をアイテム集合とする. また, $C_X \Rightarrow C_Y$ を右否定, 左否定のいずれかの形のルールとする.

S1. $X \cap Y = \emptyset$ (独立性)

S2. $\mathrm{sup}(X) \geq ms$ かつ $\mathrm{sup}(Y) \geq ms$ (前件と後件の頻出性)

S3. $\mathrm{sup}(C_X \Rightarrow C_Y) \geq ms$ (ルールの頻出性)

S4. $\mathrm{conf}(C_X \Rightarrow C_Y) \geq mc$ (ルールの確信度)

S5. $\mathrm{sup}(X \Rightarrow Y) < ms$ (無矛盾性)

無矛盾性条件は正ルール $X \Rightarrow Y$ と負ルール $C_X \Rightarrow C_Y$ が同時に抽出され, どちらが妥当かが分からなくなるような状況を回避するために導入されている.

$ms > 0.5$ の時, 負ルールの最大の支持度は 0.5 より必ず小さくなり (後の定理 1 を参照のこと), 生成される負ルールが頻出性条件 S3 を満たせず, 一つも抽出されなくなる. これは負ルールの最大支持度は $1 - ms$ であることに起因して, 無矛盾性とアイテム集合, ルールに対する最小支持度が全て同一の値に設定していることが原因と考えられる. 例として, 表 1 のデータベースでは, $ms = 0.6$ とすると, $A \Rightarrow \neg B$ と $\neg A \Rightarrow B$ は共に支持度が 0.4 であるので, 頻出性条件を満たさない. よって, 正ルールも負ルールも抽出できず, 妥当なルールは何も抽出されない. 上記の妥当性の定義では ms を 0.5 以下にしなければ負ルールは抽出されない. そのため密なデータセットには不適切な枠組となっている. この問題を解決すべく, 本研究では妥当なルールの条件を再考察する.

表 1: トランザクションデータベースの例

TID	アイテム集合
1	A
2	A
3	A, B
4	B
5	B

3.2 妥当性条件の再考察

本研究では, 3.1 節で述べた問題を解決すべく, 飽和アイテム集合の最小支持度 (以下, “msI” と略記する)

と, 正ルールと負ルールの最小支持度 (以下, “msP” “msN” と略記する) をそれぞれ別々に設定する方法を提案する.

本来, 抽出したいルールは, データセットによって変化する. 疎なデータセットならば負ルールが多く存在するため, 数少ない正ルールと, 負ルールの中で特に支持度や確信度の高いルールを抽出したい. よって msP は低く, msN は高く設定できる枠組みが欲しい. 一方, 密なデータセットならば正ルールが多く存在するため, 数少ない負ルールと, 正ルールの中で特に支持度や確信度の高いルールを抽出したい. よって msP は高く, msN は低く設定したい.

最小支持度や最小確信度を別々に設定できれば, データセットに対応して抽出したいルールを適切に抽出できると考える. しかし別々に設定する場合, 下記のような考慮すべき条件が発生する.

定理 1. X, Y を頻出アイテム集合とするとき, $msN > 1 - msI$ ならば, 任意の負ルール $C_X \Rightarrow C_Y$ に対して, $\sup(C_X \Rightarrow C_Y) < msN$ である.

証明. X と Y は頻出アイテム集合なので, $\sup(X) \geq msI$ かつ $\sup(Y) \geq msI$ である. よって負のアイテム集合 C_X と C_Y において, $\sup(\neg X) \leq 1 - msI$ かつ $\sup(\neg Y) \leq 1 - msI$ である. ここで, 図 1 から, $\sup(X \Rightarrow \neg Y) \leq \sup(\neg Y)$ となるのは明らかである. よって以下が成り立つ.

$$\sup(X \Rightarrow \neg Y) \leq 1 - msI$$

左否定, 両否定形負ルールにおいても同様に $\sup(\neg X \Rightarrow Y) \leq \sup(\neg X)$, $\sup(\neg X \Rightarrow \neg Y) \leq \sup(\neg X)$ となるため, 以下が成り立つ.

$$\sup(\neg X \Rightarrow Y) \leq 1 - msI$$
$$\sup(\neg X \Rightarrow \neg Y) \leq 1 - msI$$

以上より, 負ルールにおいて以下が成り立つ.

$$\sup(C_X \Rightarrow C_Y) \leq 1 - msI$$

これより, $msN > 1 - msI$ ならば, $\sup(C_X \Rightarrow C_Y) < msN$ となる. □

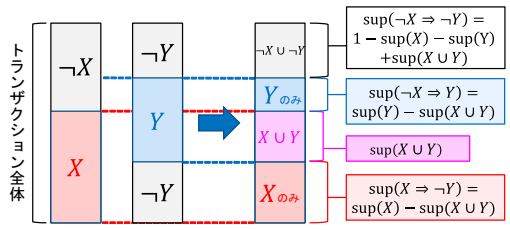

図 1: 支持度内訳

よって, 負ルールの最小支持度 msN を $1 - msI$ 以上の値に設定すると, 全ての負ルールは妥当性の条件を満たさなくなる. 2 つの最小支持度 msI と msN は, $msN \leq 1 - msI$ を満たす必要がある.

また, 同様に確信度に関しても考慮すべき条件が存在する. 最小確信度 mc を, 正ルールに対する mcP, 左否定形に対する mcL, 右否定形に対する mcR, 両否定形に対する mcD の 4 つに細分化する.

定理 2. X と Y を頻出アイテム集合とするとき mcR が, $mcR > (1 - msI)/msI$ を満たすならば, 右否定形 $X \Rightarrow \neg Y$ は常に, $\mathrm{conf}(X \Rightarrow \neg Y) < mcR$ である.

証明. 定義より,

$$\mathrm{conf}(X \Rightarrow \neg Y) = \frac{\sup(X) - \sup(X \cup Y)}{\sup(X)}$$

分子の最大値は $1 - msI$ となる (定理 1 の証明を参照のこと). よって, 以下が成り立つ.

$$\mathrm{conf}(X \Rightarrow \neg Y) \leq \frac{1 - msI}{msI}$$

以上から, $mcR > (1 - msI)/msI$ ならば, $\mathrm{conf}(X \Rightarrow \neg Y) < mcR$ となる. □

以上より, 右否定形の最小確信度 mcR を $(1 - msI)/msI$ 以上に設定すると, 全ての右否定形負ルールが抽出されなくなる. これより mcR は $mcR \leq (1 - msI)/msI$ を満たす必要がある.

以上の考察を踏まえて, **正ルール $X \Rightarrow Y$ での妥当性条件**は msI, msP と mcP を用いて, 以下の条件と定める. これらは本質的に通常の条件 [7, 11] と同じである.

V1. $X \cap Y = \emptyset$ (独立性条件)

V2. $\sup(X) \geq msI$ かつ
$\sup(Y) \geq msI$ (前件と後件の頻出性)

V3-P. $\sup(X \Rightarrow Y) \geq msP$ (ルール頻出性)

V4-P. $\mathrm{conf}(X \Rightarrow Y) \geq mcP$ (確信度)

負ルール $C_X \Rightarrow C_Y$ の妥当性条件は上述の V1 と V2 に加えて, 以下の 4 条件を加えたものと定める.

V3-N. $\sup(C_X \Rightarrow C_Y) \geq msN$ (ルール頻出性)

V4-NL. $\mathrm{conf}(\neg X \Rightarrow Y) \geq mcL$
V4-NR. $\mathrm{conf}(X \Rightarrow \neg Y) \geq mcR$
V4-ND. $\mathrm{conf}(\neg X \Rightarrow \neg Y) \geq mcD$ (確信度)

以下の V5-P は, 疎なデータセットにおいて正ルールと左右の否定形の負ルールを同時に抽出したいときに, **正ルールを優先的に抽出**させるための条件である. V5-N は, 密なデータセットにおいて左右の否定形と両否定形の負ルールを同時に抽出したいときに, **両否定形負ルールを優先的に抽出**させるための条件である.

V5-P. $\sup(X \Rightarrow Y) < msP$ (正優先無矛盾性)
V5-N. $\sup(\neg X \Rightarrow \neg Y) < msN$ (負優先無矛盾性)

また, 上記の妥当性条件の前提となる閾値には, 制約条件として,

C1. $msN \leq 1 - msI$

C2. $mcR \leq \frac{1-msI}{msI}$

を仮定するものとする.

3.3 使用する関連性尺度

先行研究 [2] では関連性尺度として, cosine 尺度を以下のように負ルールにも対応するように拡張した.

$$\mathrm{cosine}(C_X \Rightarrow C_Y) = \frac{\sup(C_X \cup C_Y)}{\sqrt{\sup(C_X)\sup(C_Y)}}$$

しかし, cosine 尺度には X と Y の出現頻度にかかわらず分子が $\sup(C_X \Rightarrow C_Y) = 0$ となれば, ルール全体の値も零となってしまう欠点がある. 例として, 表 2 のトランザクションデータベースを考える. ここでは $C \Rightarrow \neg D$ と $C \Rightarrow \neg E$ の 2 つの負ルールの cosine 値はどちらも零になる. C と D はそれぞれ 4 回, E は 2 回出現しており, E の方が出現回数が少ないことから, 本来 $C \Rightarrow \neg E$ の方が負の相関が強い. しかし, cosine 値は共に分子が零となり, 区別がつけられなくなっている. これは負ルールマイニングにおいては望ましくない現象である.

表 2: トランザクションデータベース

TID	アイテム集合
1	A, B
2	C, D, E
3	B, C, D
4	A, B
5	B, C, D
6	B, C, D, E

上述の現象は, cosine 尺度が $\sup(X \cup Y)$ と $\sup(X) \times \sup(Y)$ の比率から, ルールの前件と後件のアイテム集合間の独立性を測っていることに原因がある. よって, 本研究では $\sup(X \cup Y)$ と $\sup(X) \times \sup(Y)$ の差で独立性を測る尺度に着目する. 代表的な尺度として, 以下の **Leverage**[7] が挙げられる.

$$\mathrm{Leverage}(X, Y) = \sup(X \cup Y) - \sup(X)\sup(Y)$$

さらにこの尺度に重みをつけ, 改良したものとして, **SupportDifference**[6, 7](以降"SD" と略) と **ϕ 係数** [7] が存在する.

$$\begin{aligned}\mathrm{SD}(X \Rightarrow Y) &= P(Y|X) - P(Y|\neg X) \\ &= \frac{\sup(X \cup Y) - \sup(X)\sup(Y)}{\sup(X)(1 - \sup(X))}\end{aligned}$$

$$\phi\ (X, Y) = \frac{\sup(X \cup Y) - \sup(X)\sup(Y)}{\sqrt{\sup(X)\sup(\neg X)\sup(Y)\sup(\neg Y)}}$$

以降, それぞれの尺度の特性を考察する. Leverage と ϕ 係数には, $X \Rightarrow \neg Y$ と $\neg Y \Rightarrow X$ のような前件と後件を入れ替えたルールの区別がつけられない欠点が存在するが, これは確信度を参照することで解消できる. よって以降では, Leverage と ϕ 係数は確信度と併用することを前提として考える. (Leverage+conf, ϕ 係数+conf と略)

Leverage+conf と SD を比較する. 例として以下の表 3 のデータベースにて $A \Rightarrow B$ と $B \Rightarrow A$ の 2 つのルールについて, それぞれの尺度の値を求め, 関連性の強さを考察する.

$$\begin{aligned}&\mathrm{Lev}(A, B) = (\frac{1}{2} - \frac{2}{3} * \frac{1}{2}) = \frac{1}{6} \\ &\mathrm{conf}(A \Rightarrow B) = \frac{3}{4}, \quad \mathrm{conf}(B \Rightarrow A) = 1 \\ &\mathrm{SD}(A \Rightarrow B) = \frac{1}{6}/(\frac{2}{3} * \frac{1}{3}) = \frac{3}{4} \\ &\mathrm{SD}(B \Rightarrow A) = \frac{1}{6}/(\frac{1}{2} * \frac{1}{2}) = \frac{2}{3}\end{aligned}$$

2 つの尺度で $A \Rightarrow B$ と $B \Rightarrow A$ の関連性の強さが反対の結果となる. このうち, 関連性が強いのは, 「A が 4 回出現中で B が 3 回出現」と「B が 3 回出現中で A が 3 回出現」を考慮すれば, $B \Rightarrow A$ の方である. よって, 望ましい関連性の強さを示すのは Leverage+conf である. SD 値が逆の結果を示す原因の一つとして, SD 値の分母が $\sup(X)(1 - \sup(X))$ という形をしているため, $\sup(X)$ の値が 0.5 に近いほど SD 値は小さくなるという特性を持っていることが考えられる. SD は, 関連性の強弱を測るのに向いていないと判断できる.

次に Leverage と ϕ 係数を比較する. 以下の表 4 にて $A \Rightarrow B$ と $C \Rightarrow D$ の 2 つのルールについて, それぞれの尺度の値を求め統計的な意味での相関性の強さを測る.

$$\begin{aligned}&\mathrm{Lev}(A, B) = (\frac{2}{6} - (\frac{2}{6} * \frac{2}{6})) = \frac{8}{36} \\ &\mathrm{Lev}(C, D) = (\frac{3}{6} - (\frac{3}{6} * \frac{3}{6})) = \frac{1}{4} \\ &\phi\ (A, B) = \frac{8}{36}/\sqrt{(\frac{2}{6} * \frac{2}{6} * \frac{4}{6} * \frac{4}{6})} = 1 \\ &\phi\ (C, D) = \frac{1}{4}/\sqrt{(\frac{3}{6} * \frac{3}{6} * \frac{3}{6} * \frac{3}{6})} = 1\end{aligned}$$

$A \Rightarrow B$ と $C \Rightarrow D$ の 2 つのルールは, それぞれ出現アイテム数は異なるが相関性の強さは同じである. しかし, Leverage は出現アイテム数に応じて値が変化してしまう. これに対して ϕ 係数では値が変わらず相関性が同じことを示している. これより本論文では, ϕ 係数+conf を関連性尺度として用いる.

表 3: Lev+conf と SD の比較用データベース

TID	アイテム集合
1	A, B
2	A, B
3	A, B
4	A
5	C
6	C

表 4: Lev と ϕ の比較用データベース

TID	アイテム集合
1	A, B
2	A, B
3	C, D
4	C, D
5	C, D
6	E

4 分枝限定法を用いた上位 k ルール抽出アルゴリズム

本論文では第 4 章で述べた妥当なルール条件と関連性尺度を考慮したうえで, 正負のルールを対象とした効率的な相関ルール抽出手法を提案する.

4.1 分枝限定法のための上界関数

本論文では正負のルール抽出に分枝限定法を用いる. ある頂点が表現するルールを評価する尺度の上界が閾値を満たさない時, その子節点を枝刈りすることで, 探索空間の削減を行える. 分枝限定法の枝刈りのために, ϕ 係数, 左否定形の確信度の上界関数として以下を考える.

$$\overline{\mathrm{conf}}(\neg X \Rightarrow Y) = \frac{\sup(Y)}{1-\sup(X)}$$

$$\overline{\phi_{\mathrm{PR}}}(X,Y) = \frac{\sup(Y)-\sup(X)\sup(Y)}{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y))}}$$

2 引数関数 $f(X,Y)$ が**右逆単調性**であるとは, $Y \leq Y'$ なる Y と Y' に対して $f(X,Y) \leq f(X,Y')$ を満たす場合を言う. 以下で説明するように $\overline{\phi_{\mathrm{PR}}}$ は右逆単調性を満たす ϕ の上界関数である[1]. 一方 $\overline{\mathrm{conf}}$ は X と Y の双方に関して逆単調である.

また, ϕ 係数は正と負の相関の両方を示せるが, 枝刈のためには, 正ルールの判定には上界を, 負ルールの判定には下界を考える必要がある. よって, ϕ 係数で負の相関を計るために, 以下の下界関数を考える.

$$\underline{\phi_{\mathrm{N}}}(X,Y) = \frac{-\sup(X)\sup(Y)}{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y))}}$$

[1]次の $\overline{\phi_{\mathrm{PL}}}$ は左逆単調性を満たす上界関数であるが, 本論文では前件を固定させ, 後件を変化させる探索を行うため, 使用しない.

$$\overline{\phi_{\mathrm{PL}}}(X,Y) = \frac{\sup(X)-\sup(X)\sup(Y)}{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y))}}$$

$\underline{\phi_{\mathrm{N}}}$ は X と Y 双方に対して単調性を満たす, 下界関数である.

定理 3. 関数 $\overline{\mathrm{conf}}$ は関数 conf の上界をなし, X, Y 双方に対して逆単調性を満たす.

証明. $\sup(X \cup Y) \geq 0$ より, 以下が明らかに成り立つ.

$$\begin{aligned}\mathrm{conf}(\neg X \Rightarrow Y) &= \frac{\sup(Y)-\sup(X \cup Y)}{1-\sup(X)} \\ &\leq \frac{\sup(Y)}{1-\sup(X)} = \overline{\mathrm{conf}}(\neg X \Rightarrow Y)\end{aligned}$$

よって, $\overline{\mathrm{conf}}$ は conf の上界関数となる. また, $X \subset X'$ なるアイテム集合 X と X' に対して, $\sup(X) \geq \sup(X')$ なので, 以下は明らかである.

$$\frac{\sup(Y)}{1-\sup(X)} \geq \frac{\sup(Y)}{1-\sup(X')}$$

同様に, $Y \subset Y'$ なるアイテム集合 Y と Y' に対しても, 以下は明らかである.

$$\frac{\sup(Y)}{1-\sup(X)} \geq \frac{\sup(Y')}{1-\sup(X)}$$

よって, X, Y 双方に対して, 逆単調性が成り立つ. □

定理 4. 関数 $\overline{\phi_{\mathrm{PR}}}$ は関数 ϕ の上界をなし, Y に対して逆単調性を満たす.

証明. $\sup(X \cup Y) \leq \sup(Y)$ より以下が明らかに成り立つ.

$$\begin{aligned}\phi\ (X,Y) &= \frac{\sup(X \cup Y)-\sup(X)\sup(Y)}{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y))}} \\ &\leq \frac{\sup(Y)-\sup(X)\sup(Y)}{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y))}} = \overline{\phi_{\mathrm{PR}}}(X,Y)\end{aligned}$$

よって, $\overline{\phi_{\mathrm{PR}}}$ は ϕ の上界となる.

また, 右逆単調性は, $Y' \subseteq Y$ を満たすアイテム集合 Y と Y' に対して, $\overline{\phi_{\mathrm{PR}}}(X,Y) - \overline{\phi_{\mathrm{PR}}}(X,Y') \geq 0$ を示すことで証明する. 支持度の差 Δ を $\Delta = \sup(Y) - \sup(Y')$ とする.

$$\begin{aligned}&\overline{\phi_{\mathrm{PR}}}(X,Y) - \overline{\phi_{\mathrm{PR}}}(X,Y') = \\ &\quad\frac{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y))(1-\sup(Y'))}}{(1-\sup(Y))(1-\sup(Y'))\sup(X)} - \\ &\quad\frac{\sqrt{\sup(X)(1-\sup(X))\sup(Y')(1-\sup(Y'))(1-\sup(Y))}}{(1-\sup(Y))(1-\sup(Y'))\sup(X)}\end{aligned}$$

この式について, $\sup(Y) = a$, $\sup(X) = b$, $\sup(Y') = a - \Delta$ とおくと以下のように変形できる.

$$\begin{aligned}&\overline{\phi_{\mathrm{PR}}}(X,Y) - \overline{\phi_{\mathrm{PR}}}(X,Y') = \\ &\quad\frac{\sqrt{a(1-a)b(1-b)}(1-a+\Delta)}{(1-a)(1-a+\Delta))b} - \\ &\quad\frac{\sqrt{(a-\Delta)(1-a+\Delta)b(1-b)}(1-a)}{(1-a)(1-a+\Delta))b}\end{aligned}$$

$0 \leq a, b, \Delta \leq 1$ より, 分母は正である. よって分子が正であることを示せばよい. 分子の部分を F としたとき, 各項の共通部分である $\sqrt{b(1-b)}$ を, 簡略化のため

に削除できる．$\sqrt{b(1-b)} \geq 0$ であることに注意する．簡略した F を $F(a, \Delta)$ と表記するとき，

$$F(a, \Delta) = \sqrt{a(1-a)}(1-a+\Delta) - \sqrt{(a-\Delta)(1-a+\Delta)}(1-a)$$

a, Δ 共に正の値をとるため，以下の**二乗比較**の定理を用いる．

- $\alpha > 0$, $\beta > 0$ の時，$\alpha^2 - \beta^2 \geq 0$ ならば $\alpha - \beta \geq 0$.

$F(a, \Delta)$ の各項を二乗し，まとめると以下のようになる．

$$\begin{aligned} F(a, \Delta)^2 &= a^2\Delta - a\,\Delta^2 - 2a\,\Delta + \Delta^2 + \Delta \\ &= ((a-1)\,\Delta\,(a-\Delta-1)) \end{aligned}$$

$0 \leq a, b, \Delta \leq 1$ より，$(a-1)$ は負，Δ は正，$(a-\Delta-1)$ は負である．よって $F(a, \Delta)^2 \geq 0$ から，$F(a, \Delta) \geq 0$ を満たし，分子も正となることが示される．以上より，$\underline{\phi}_{\mathrm{PR}}$ は右逆単調性を満たすことが保証される． □

定理 5. 関数 $\underline{\phi}_{\mathrm{N}}$ は，関数 ϕ の下界をなし，X, Y 双方に対して単調性を満たす．

証明. $\underline{\phi}_{\mathrm{N}}$ が ϕ の下界となることは式の形より明らかである．単調性は，$X \subseteq X'$ を満たすアイテム集合 X と X' に対して，$\underline{\phi}_{\mathrm{N}}(X, Y) - \underline{\phi}_{\mathrm{N}}(X', Y) \leq 0$ を示すことで証明する．支持度の差 Δ を $\Delta = \sup(X) - \sup(X')$ とする．

$$\begin{aligned} &\underline{\phi}_{\mathrm{N}}(X, Y) - \underline{\phi}_{\mathrm{N}}(X', Y) = \\ &\quad \frac{\sqrt{\sup(X')(1-\sup(X'))\sup(Y)(1-\sup(Y)}(1-\sup(X))}{(1-\sup(X))(1-\sup(X'))(1-\sup(Y))} - \\ &\quad \frac{\sqrt{\sup(X)(1-\sup(X))\sup(Y)(1-\sup(Y)}(1-\sup(X'))}{(1-\sup(X))(1-\sup(X'))(1-\sup(Y))} \end{aligned}$$

この式について，$\sup(X) = a$, $\sup(Y) = b$, $\sup(X') = a - \Delta$ とおくと以下のように変形できる．

$$\begin{aligned} &\underline{\phi}_{\mathrm{N}}(X, Y) - \underline{\phi}_{\mathrm{N}}(X', Y) = \\ &\quad \frac{\sqrt{(a-\Delta)(1-a+\Delta)b(1-b)}(1-a)}{(1-a)(1-a+\Delta))(1-b)} - \\ &\quad \frac{\sqrt{a(1-a)b(1-b)}(1-a+\Delta)}{(1-a)(1-a+\Delta))(1-b)} \end{aligned}$$

$0 \leq a, b, \Delta \leq 1$ より，分母は正である．よって分子が負であることを示せばよい．分子の部分を F としたとき，各項の共通部分である $\sqrt{b(1-b)}$ を含む部分を，簡略化のために削除できる．$\sqrt{b(1-b)} \geq 0$ であることに注意する．簡略した F を $F(a, \Delta)$ と表記するとき，

$$F(a, \Delta) = \sqrt{(a-\Delta)(1-a+\Delta)}(1-a) - \sqrt{a(1-a)}(1-a+\Delta)$$

a, Δ 共に正の値をとるため，前証明で用いた二乗比較の定理を用いる．$F(a, \Delta)$ の各項を二乗し、まとめると以下のようになる．

$$\begin{aligned} F(a, \Delta)^2 &= -a^2\Delta + a\,\Delta^2 + 2a\,\Delta - \Delta^2 - \Delta \\ &= -((a-1)\,\Delta\,(a-\Delta-1)) \end{aligned}$$

$0 \leq a, b, \Delta \leq 1$ より，$(a-1)$ は負，Δ は正，$(a-\Delta-1)$ は負である．よって $F(a, \Delta)^2 \leq 0$ から，$F(a, \Delta) \leq 0$ を満たし，分子は負となることが示される．以上より，$\underline{\phi}_{\mathrm{N}}$ は単調性を満たすことが保証される． □

4.2 上位 k ルール抽出擬似コード

本論文で提案するアルゴリズムの擬似コードを Alg.1 に挙げる．本論文では，疎なデータに対して正ルールと左右の否定形ルールを抽出するコードを紹介する．即ち正ルールの妥当性 V1, V2, V3-P, V4-P と，ならび負ルールの妥当性 V1, V2, V3-N, V4-NL, V4-NR, および V5-P の条件を満たす，正と負のルールを抽出するコードである．このコードは関連性尺度に関する上位 k 個のルールを抽出する．本論文では先行研究 [1, 7] に基づき，接尾辞木を用いた探索を行う．関連性尺度の閾値は，正ルールに対する mrP，左否定形負ルールに対する mrL，右否定形負ルールに対する mrR，両否定形負ルールに対する mrD の 4 つに細分化する．擬似コード中の 10〜12 行目で X と Y の重複性 (条件 V1) を確認する．14〜24 行目で正ルールの候補が妥当性を満たすか否かを判定する．25〜35 行目では左否定形，36〜46 行目で右否定形の候補ルールが妥当性を満たすか否かを判定する．47〜49 行目では，ルール候補が全ての上界関数を満たさない場合に子節点を枝刈りする．

5 実験と考察

本研究で提案したアルゴリズムの効率性を 2 つの実証実験にて検証した．実験には，Frequent Itemset Mining Dataset Repository[9] から 3 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 5 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．また，論文では時間の都合上，密なデータセットで両否定形を抽出する実験を実施していない．今後の課題である．

表 5: 実験に使用したデータセット

データセット	#(item)	#(trans)	ave(item)
mushroom	119	8,124	23
retail	16,470	88,162	10.3
connect	130	67,557	43

5.1 分枝限定法の有効性の実験

上界関数を用いた枝刈の効果を，ルール候補 $X \Rightarrow Y$ の前件 X と後件 Y の対 (以下，**検査対**と略) の数と探索時間，枝刈された回数で比較する．実験の際に用いたパラメータ条件を以下の表 6 に示す．これらのパラメータは，C1 と C2 の制約条件を満たすことに注意されたし．

Algorithm 1 疎なデータに対する上位 k 抽出アルゴリズム

Input: データセット D から抽出した極小生成子の集合 MGS, 抽出ルール数 k, 最小支持度 (正)msP, (左・右否定)msN, 最小確信度 (正)mcP, (左否定)mcL, (右否定)mcR
Output: 正, 左・右否定形の上位 k ルールリスト

```
==========
mrP, mrL, mrR ← 0;                          ▷ 関連性尺度の暫定閾値
pn, ln, rn ← 0                              ▷ 各上位 k リストの合計ルール数
FP, FL, FR ←False;                          ▷ 上界関数を満たすかのフラグ
PL, LL, RL ←nil;                            ▷ 各上位 k リストの初期化
MGS から接尾辞木を作成し,
MGS の極小生成子を左優先深さ優先で並べる
その結果を MG_1, MG_2, ..., MG_N なる列とする

for i ← 1 to N do
    X ← MG_i;
    for j ← 1 to N do
        Y ← MG_j;
        if X ∩ Y ≠ ∅ then                   ▷ 独立性検査
            Y の子節点を枝狩り;
        else
            FL, FR, FP←False    ▷ 上界関数のフラグ初期化
    ▷ [正ルール (X ⇒ Y) の上界関数を満たすか判定]======
            if sup(X⇒Y)≥msP and φ_PR(X,Y)≥mrP then
                FP ←True;
                if conf(X ⇒ Y) ≥ mcP and
                   φ (X,Y) ≥ mrP then
                    PL にルールを φ 値降順で追加; pn ← pn+1;
                    if pn > k then
                        PL の末尾を削除;
                        mrP を削除後の末尾の φ 値に更新;
                    end if
                end if
            end if
    ▷ [左否定形 (¬X ⇒ Y) の上界関数を満たすか判定]======
            if conf(¬X⇒Y)≥mcL and φ_N(X,Y)≤mrL then
                FL ←True;
                if sup(¬X⇒Y)≥msN and sup(X⇒Y)≤msP
                conf(¬X⇒Y)≥mcL and φ(X,Y)≤mrL then
                    LL にルールを φ 値昇順で追加; ln ← ln+1;
                    if ln > k then
                        LL の末尾を削除;
                        mrL を削除後の末尾の φ 値に更新;
                    end if
                end if
            end if
    ▷ [右否定形 (X ⇒ ¬Y) の上界関数を満たすか判定]======
            if φ_N(X,Y)≤mrR then
                FR ←True;
                if sup(X⇒¬Y)≥msN and sup(X⇒Y)≤msP
                conf(X⇒¬Y)≥mcR and φ(X,Y)≤mrR then
                    RL にルールを φ 値昇順で追加; rn ← rn+1;
                    if rn > k then
                        RL の末尾を削除;
                        mrR を削除後の末尾の φ 値に更新;
                    end if
                end if
            end if
            if (FL ∧ FR ∧ FP) = False then
                上界関数より, Y の子節点を枝狩り
            end if
        end if
    end for
end for
```

また, Alg.1 を用いて関連性尺度でそれぞれ上位 100 個の正ルール, 左否定形, 右否定形のルールを抽出した結果を表 7 に示す. 表の枝刈回数欄にある「重複」は, 重複性条件 (条件 $V1$) による枝刈回数, 「正 $\overline{\phi}$」・「左 $\underline{\phi}$」・「右 $\underline{\phi}$」はそれぞれ ϕ 係数の上界または下界関数による枝刈回数, 「正 sup」は正ルール支持度による枝刈回数, 「左 $\overline{\text{conf}}$」は左否定形確信度上界関数による枝刈回数を示している.

表 6: パラメータ条件

case	mcL	mcR	k	msI	msP	msN
$M1$	0.5	0.5	100	0.1	0.1	0.1
$M2$				0.2	0.2	0.2
$M3$				0.3	0.3	0.3
$M4$				0.4	0.4	0.4
$R1$	0.5	0.5	100	0.001	0.001	0.1
$R2$				0.002	0.002	0.2
$R3$				0.003	0.003	0.3
$C1$	0.5	0.1	100	0.8	0.81	0.1
$C2$				0.9	0.91	0.1

結果より, mushroom, retail にて検査対・探索時間が削減されていることが確認できた. しかし, connect では正の支持度による枝刈はできているが, 検査対の数は全く変化していない. これは, 子節点の枝刈は正ルール, 左右の負ルールの探索が全て必要ない時にだけ行うことができるため, 正の支持度条件により正ルールだけの枝刈が可能になっても, 子節点の枝刈ができないためである. connect で ϕ 上界関数が機能していないのは, connect では各ルール間の ϕ 係数の値に殆ど差がないためである. 一方, mushroom では $msI = 0.3$ の時, 枝刈しないと 13 万個の検査対に対して 3.54 秒かかるが, 枝刈して残った 7 万の検査対に対しては 3.31 秒かかっている. これより, 枝刈処理にオーバーヘッドが存在していることが確認できた.

5.2 関連性尺度の比較

関連性尺度 ϕ を用いる本手法と Leverage, SD, cosine それぞれを用いる手法で抽出されたルールの傾向を比較・考察する. mushroom, retail, connect の全てで表 6 の条件を用いて抽出ルール数 k =100 として, Alg.1 の実装コードを用いて実験を行った.

結果として, Leverage は $\sup(X)$ と $\sup(Y)$ それぞれが 0.5 に近く, $\sup(X \cup Y)$ が 1 に近い正ルール, 0 に近い負ルールを抽出する傾向であることを確認した. 一方で, SD, ϕ 係数は $\sup(X)$ と $\sup(Y)$ の和が大きく, 差が小さく, $\sup(X \cup Y)$ が 1 に近い正ルール, 0 に近い負ルールを抽出する傾向であることを確認した. さらに, SD は $\sup(X)$ が 0.5 に近いルールを優先し, 正ルールにおいて関連性の強さの順序が逆になっている部分があることを確認した. また, cosine は ϕ 係数と傾向が共通していることを確認した. 以上より, 各関連性尺度を用いたときの抽出ルールは概ね 3.3 節で述べた予測と一致することが確認できた.

表 7: 関連性尺度で上位 100 個の正負のルールの分枝限定法による抽出実験

データ	条件	sup 検査対		探索	枝刈回数					
				時間 [s]	重複	正 $\overline{\phi}$	正 sup	左 $\underline{\phi}$	左 $\overline{\text{conf}}$	右 $\underline{\phi}$
mushroom	$M1$	×	28,634,310	179.37	4,523,234	0	0	0	0	0
		○	5,063,960	29.82	858,033	2,037,798	1,923,382	4,953,169	925	4,953,660
	$M2$	×	1,499,788	9.34	264,510	0	0	0	0	0
		○	457,822	5.78	96,030	175,827	158,872	417,993	155	418,007
	$M3$	×	136,210	3.54	33,012	0	0	0	0	0
		○	71,256	3.31	13,314	13,083	47,945	51,693	0	51,693
	$M4$	×	16,136	2.83	3,628	0	0	0	0	0
		○	14,681	2.81	3,086	753	12,617	4,172	0	4,414
retail	$R1$	×	42,398,816	62.20	9,375,458	0	0	0	0	0
		○	16,997,316	42.04	76,582	63,506	16,910,128	7,478,993	9,513,533	16,675,555
	$R2$	×	5,640,032	13.99	1,043,172	0	0	0	0	0
		○	4,487,003	13.36	506,500	0	4,480,229	4,012	4,481,214	2,863,293
	$R3$	×	1,510,850	6.89	272,543	0	0	0	0	0
		○	1,510,850	6.83	272,543	0	1,507,434	0	1,509,923	0
connect	$C1$	×	66,183,208	52.16	21,303,058	0	0	0	0	0
		○	66,183,208	52.16	21,303,058	0	65,674,928	0	0	0
	$C2$	×	3,159,824	4.38	1,308,476	0	0	0	0	0
		○	3,159,824	4.36	1,308,476	0	3,086,330	0	0	0

6 まとめ

本研究では, 正負の相関ルールの妥当性を再考察し, それを踏まえた上での関連性尺度を用いた効率的な正負の相関ルール抽出手法を提案した. また, ルール抽出に分枝限定法を適用するために幾つかの評価尺度の逆単調な上界関数を与え, その有効性を実証実験にて確認した.

今後の課題として, まだ確認していない密なデータセットにおける両否定形負ルールの抽出実験, または ϕ 係数上界関数の再考察がある.

謝辞

本研究は, ISPS 科学研究費補助金（No.16K00298）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] 黒岩健歩, 岩沼宏治, 山本泰生: 負相関ルールマイニングの高速化と関連性尺度の導入, 人工知能学会第 97 回人工知能基本問題研究会, SIG-FPAI, B4-04, pp. 7-12 (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ルール抽出計算の効率化, 人工知能学会全国大会第 31 回, pp. 1-4 (2017)

[4] 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)

[5] Mohammed, J. Z.: Mining Non-Redundant As sociation Rules. Data Mining and Knowledge Discovery, 9, pp. 223-248 (2004)

[6] Bay, S. D. and Pazzani, M. J.: Detecting group differences: mining contrast sets, Data Mining and Knowledge Discovery, Vol. 5, pp. 213-246 (2001)

[7] 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)

[8] Han, J., Wang, J., Lu, Y. and Tzvetkov, P.:Mining top- K frequent closed patterns without minimum support, In Proc. of the 2002 IEEE Int'l Conf. on Data Mining(ICDM-02), pp. 211-218 (2002).

[9] Frequent Itemset Mining Dataset Repository, <http://fimi.ua.ac.be/>(2018-2-9)

[10] LCM: Linear time Closed itemset Miner, <http://research.nii.ac.jp/ uno/code/lcm.html> (2018-2-9)

[11] J. Han, J. Pei and Y. Yin: Frequent Patterns without Candidate Gerneration. *Proc. ACM-SIGMOD'00,* pp1-12, (2000)

[12] X. Wu, C. Zhang, and S. Zhang: Efficient Mining of Both Positive and Negative Association Rules. *ACM-Trans. on Information Systems,* Vol.22(3), pp.381-405, (2004)

[13] H. Wang, X. Zhang and G. Chen: Mining a Complete Set of Both Positive and Negative Association Rules from Large Databases. *Proc. the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD'08),* pp.777-784, (2008)