不確実データベースからの負の相関ルールの発見

Discovery of Negative Association Rules from Uncertain Databases

藤田 岳行 [1*]　尾崎 知伸 [1]
Takeyuki Fujita[1] and Tomonobu Ozaki[1]

[1] 日本大学 文理学部
[1] College of Humanities and Sciences, Nihon University

Abstract: Uncertain data mining attracts much attention recently. In this paper, as an extension of probabilistic association rules, we propose probabilistic negative association rules to capture strong relationships between presence and absence of itemsets in uncertain databases. An algorithm is developed for extracting probabilistic negative association rules, which combines probabilistic frequent itemsets stored in a postfix tree.

1　はじめに

近年，ソーシャルネットワークの広がりやセンサネットワークの発展により，大規模なデータが容易に蓄積されるようになった．その一方で，個人情報の保護を目的に，生のデータにダミーデータなどのノイズを加えたり，集約を行うといった操作が行われるようになっている．また，センサーネットワークから得られるデータには，ノイズが含まれる可能性も大きい．こうした背景のもと，不確実なデータを対象とした分析手法が盛んに研究されている [1, 2].

不確実データを対象とした分析技術の一つとして，確率的データベースから頻出パターンや相関ルールを発見する手法が提案されている [11, 3, 12]．相関ルール分析とは，主に，コンビニエンスストアに代表される小売業におけるバスケット分析などで利用されている分析手法である．データベースの中で同時に発生することの多い，すなわち相関の高いアイテム集合（商品集合）X と Y の対を，ルール $X \Rightarrow Y$ として抽出する．

一般に相関ルール分析では，2 つのアイテム集合の出現に着目するが，「ある商品を購入した場合は，別の商品は購入しない」など，負の出現も大きな意味を持つ場合も考えられる．この考えに従い，近年，$X \Rightarrow \neg Y$ や $\neg X \Rightarrow Y$，$\neg X \Rightarrow \neg Y$ の形式で表現される負の相関ルールに関して，多くの研究が行われている [14, 5, 13].

これらの背景のもと，本研究では，今後ますますの増大が予想される確率的データベースを対象に，負の相関ルールを抽出するアルゴリズムを提案する．提案するアルゴリズムは，分割統治法に基づく効率的な確率計算 [11] を利用して確率的頻出パターンを求め，それらを接尾辞木に配置 [7, 8, 9] した上で組み合わせを考慮することで，ルールを発見するものである．

本論文の構成は以下のとおりである．2 章と 3 章で，既存研究に基づき，負の相関ルール及び確率的相関ルールを導入する．次いで，4 章で確率的負の相関ルールを提案し，その発見アルゴリズムを示す．5 章で実験結果を報告し，最後に 6 章でまとめと今後の課題を述べる．

2　負の相関ルール

$I = \{a_1, a_2, \cdots, a_n\}$ をアイテムの集合とする．このとき $T \subseteq I$ となるアイテムの部分集合 T をトランザクションと呼ぶ．また，トランザクションの集合 $DB = \{T_1, T_2, \cdots, T_k\}$ をデータベースと呼ぶ．

$X \subseteq I$, $Y \subseteq I$, $X \cap Y = \emptyset$ であるアイテム集合 X と Y に対し，$X \Rightarrow Y$ の形式をしたルールを相関ルールと呼ぶ．このとき，X をルールの前件，Y をルールの後件と呼ぶ．

データベース DB におけるアイテム集合 X の支持度 $sup(X)$ とは，X の出現回数，すなわち

$$sup(X) = |\{T \in DB \,|\, X \subseteq T\}|$$

である．また，相関ルール $X \Rightarrow Y$ の支持度 $sup(X \Rightarrow Y)$ と確信度 $conf(X \Rightarrow Y)$ は，それぞれ，以下のように定義される．

$$sup(X \Rightarrow Y) = sup(XY)$$

$$conf(X \Rightarrow Y) = \frac{sup(XY)}{sup(X)}$$

なお本論文では簡略化のため，$X \cup Y$ を XY と表記する．またこれは，X と Y の両方の条件を満たすと解釈する．

*連絡先：日本大学 文理学部
〒 156-8550 世田谷区桜上水 3-25-40

アイテム集合 $X = \{a_1, a_2, \cdots, a_{|X|}\}$ の否定として，負のアイテム集合を $\neg X = \neg(a_1 \wedge a_2 \wedge \cdots \wedge a_{|X|})$ と定義する [5]．このとき，負の相関ルールとは，負のアイテム集合をルールの前件もしくは後件にとる相関ルールを指し，$X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$ となる任意のアイテム集合 X と Y について，(1)$X \Rightarrow \neg Y$，(2)$\neg X \Rightarrow Y$ 及び (3)$\neg X \Rightarrow \neg Y$ の 3 種を考えることができる．本論文ではこれらをまとめて，$C_X \Rightarrow C_Y$ と表記する．

既存研究 [14] に従い，負の相関ルール $C_X \Rightarrow C_Y$ の支持度と確信度を，以下のように定義する．

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

ここで，N はデータベース内のトランザクション数，$C_X \in \{X, \neg X\}, C_Y \in \{Y, \neg Y\}$ である．また，$\neg(C_X = X \wedge C_Y = Y)$ とする．

正の相関ルール同様，負の相関ルールの発見においても，ユーザによる 2 つの閾値，最小支持度 $\mathtt{ms}\,(0 < \mathtt{ms} \leq N)$ と最小確信度 $\mathtt{mc}\,(0 < \mathtt{mc} \leq 1)$ を満たすルールを有益なルールと判断する．しかし，正の相関ルールに比べ，負の相関ルールではその数が大きくなることが懸念される．従って，既存手法 [5, 7] に従い，抽出対象をより特徴的かつ興味深いルールに絞り込む．具体的には，$sup(C_X \Rightarrow C_Y) \geq \mathtt{ms}$ かつ $conf(C_X \Rightarrow C_Y) \geq \mathtt{mc}$ であり，さらに以下のすべての条件を満たすルール $C_X \Rightarrow C_Y$ を有効な負の相関ルールと呼ぶ．

1. $X \cap Y = \emptyset$
2. $sup(X) \geq \mathtt{ms} \ \wedge \ sup(Y) \geq \mathtt{ms}$
3. $sup(XY) < \mathtt{ms}$
4. $conf(C_X \Rightarrow C_Y) \geq \mathtt{mc}$

3 確率的相関ルール

不確実データベースとは，各トランザクションまたはトランザクション中の各アイテムに，その存在確率が付与されているデータの集合である [3]．本研究では，各アイテムに対して存在確率が付与される不確実データベースを対象とする．

トランザクション中のアイテム i_n の存在確率を p_n と表記する．このとき，トランザクション $T = \{i_1, \cdots, i_{|T|}\}$ 中の各アイテムに対し，その存在確率を付与したトランザクション $T' = \{i_1 : p_1, \ldots, i_{|T|} : p_{|T|}\}$ を不確実トランザクションと呼ぶ．また，不確実トランザクションの集合 $UDB = \{T'_1, T'_2, \cdots, T'_{|UDB|}\}$ を不確実データベースと呼ぶ．

多くの既存研究同様，本研究でも，各アイテムの存在確率は独立であると仮定した上で，可能世界意味論 [6] に従って不確実データベースを解釈する．これにより，アイテム集合及び相関ルールの支持度は確率変数となり，その分布を考えることで，最小支持度などの条件を満たす確率を求めることが可能となる．

今，$W = \{w_1, w_2, \cdots, w_n\}$ を可能世界 w_j の全体集合とする．このとき，[11] に従い，アイテム集合 X の支持度が i となる確率を以下のように表記する．

$$
P_i(X) = \sum_{w_j \in W, (S(X, w_j) = i)} P(w_j)
$$

ここで $S(X, w_j)$ は可能世界 w_j における X の支持度，$P(w_j)$ は w_j の存在確率を，それぞれ表す．

同様に，アイテム集合 X の支持度が i 以上となる確率を

$$
P_{\geq i}(X) = \sum_{w_j \in W, (S(X, w_j) \geq i)} P(w_j)
$$

と表記する．ユーザによる 2 つの閾値，最小支持度 $\mathtt{ms}$ と最小確率 $\mathtt{mp}\,(0 < \mathtt{mp} \leq 1)$ が与えられたとき，$P_{\geq \mathtt{ms}}(X) \geq \mathtt{mp}$ を満たす，すなわち「支持度が $\mathtt{ms}$ 以上となる確率が $\mathtt{mp}$ 以上」となるアイテム集合 X を，確率的頻出アイテム集合と呼ぶ [11]．

最小支持度 $\mathtt{ms}$ と最小確信度 $\mathtt{mc}$ に対し，(正の) 相関ルール $X \Rightarrow Y$ が両者を満たす確率を

$$
P(X \Rightarrow Y) = P\left[sup(XY) \geq \mathtt{ms} \wedge conf(XY) \geq \mathtt{mc}\right]
$$

と表記する．このとき，$P(X \Rightarrow Y) \geq \mathtt{mp}$，すなわち確率 $P(X \Rightarrow Y)$ が最小確率 $\mathtt{mp}$ 以上の相関ルール $X \Rightarrow Y$ を確率的頻出相関ルールと呼ぶ．

なお，確率 $P(X \Rightarrow Y)$ は，

$$
\begin{aligned}
&P\left[sup(XY) \geq \mathtt{ms} \wedge conf(XY) \geq \mathtt{mc}\right] \\
&\qquad = \sum_{i=\mathtt{ms}}^{N} P_i(XY) \sum_{j=0}^{\frac{(1-\mathtt{mc}) \cdot i}{\mathtt{mc}}} P_j(X \neg Y)
\end{aligned}
$$

と変形するとともに，$P_a(X)$ と $P_b(XY)$ から $P_c(X \neg Y)$ $(0 \leq a, b, c \leq N)$ を導出することで，不確実データベースを可能世界へと展開することなく，計算することが可能である [11]．

4　確率的負の相関ルールの発見

4.1　確率的負の相関ルール

不確実データベース上で負の相関ルールを解釈することで，確率的負の相関ルールを提案する．

最小支持度 ms と最小確信度 mc が与えられたとき，$X \cap Y = \emptyset$ なる $C_X \in \{X, \neg X\}$ と $C_Y \in \{Y, \neg Y\}$ から構成される負の相関ルール $C_X \Rightarrow C_Y$ ($\neg(C_X = X \wedge C_Y = Y)$) が，有効な負の相関ルールとなる確率を以下のように表記する．

$$P(C_X \Rightarrow C_Y) = P\left[\begin{array}{l} sup(C_X \Rightarrow C_Y) \geq \mathtt{ms} \wedge \\ conf(C_X \Rightarrow C_Y) \geq \mathtt{mc} \wedge \\ sup(X) \geq \mathtt{ms} \wedge \\ sup(Y) \geq \mathtt{ms} \wedge \\ sup(XY) < \mathtt{ms} \end{array}\right]$$

このとき，$P(C_X \Rightarrow C_Y) \geq \mathtt{mp}$，すなわち最小確率 mp を満たすルール $C_X \Rightarrow C_Y$ を，有効な確率的負の相関ルールと定義し，本研究での抽出対象とする．

また，より意味のあるルールに絞り込むため，通常の負の相関ルール同様 [5, 7]，抽出対象を極小なルールに限定する．ここで，ルール $C_X \Rightarrow C_Y$ が極小とは，$C'_X \subseteq C_X$ かつ $C'_Y \subseteq C_Y$ かつ $\neg(C'_X = C_X \wedge C'_Y = C_Y)$ を満たす，有効な確率的負の相関ルール $C'_X \Rightarrow C'_Y$ が存在しないことを意味する．

確率 $P(C_X \Rightarrow C_Y)$ に対し，(正の) 確率的相関ルールと同様の変形 [11] を施す．これにより，不確実データベースを可能世界へと展開することなく，$P(C_X \Rightarrow C_Y)$ の計算が可能となる．

$$P(X \Rightarrow \neg Y) = \sum_{i=\mathtt{ms}}^{N} P_i(X\neg Y) \sum_{j=0}^{\min\left(\mathtt{ms}-1, \frac{(1-\mathtt{mc})\cdot i}{\mathtt{mc}}\right)} P_j(XY) \sum_{k=\mathtt{ms}-j}^{N} P_k(\neg XY)$$

$$P(\neg X \Rightarrow Y) = \sum_{i=\mathtt{ms}}^{N} P_i(\neg XY) \sum_{j=0}^{\min\left(\frac{(1-\mathtt{mc})\cdot i}{\mathtt{mc}}, N-\mathtt{ms}-i\right)} P_j(\neg X\neg Y) \sum_{k=0}^{\mathtt{ms}-1} P_k(XY)$$

$$P(\neg X \Rightarrow \neg Y) = \sum_{i=\mathtt{ms}}^{N} P_i(\neg X\neg Y) \sum_{j=0}^{\min\left(\frac{(1-\mathtt{mc})\cdot i}{\mathtt{mc}}, N-\mathtt{ms}-i\right)} P_j(\neg XY) \sum_{k=0}^{N-\mathtt{ms}-i} P_k(X\neg Y) \sum_{l=0}^{\mathtt{ms}-1} P_l(XY)$$

具体的な変形過程に関しては，3 種のルールともに類似しているので，$P(X \Rightarrow \neg Y)$ に関してのみ図 1 に示し，残りの 2 つ $P(\neg X \Rightarrow Y)$ 及び $P(\neg X \Rightarrow \neg Y)$ に関しては省略する．

4.2　確率的負の相関ルール抽出アルゴリズム

本研究では，(確率的ではない) 負の相関ルール抽出に関する既存手法 [7] を確率的データベースへと拡張し，

1. 確率的頻出アイテム集合を導出し，接尾辞木に配置する
2. 接尾辞木上の各アイテム集合同士を組み合わせる

という手順で，確率的負の相関ルールを抽出するアルゴリズムを提案する．

なお，負の相関ルール $C_X \Rightarrow C_Y$ が，有効な確率的負の相関ルールとなるためには，少なくとも条件 $P_{\geq \mathtt{ms}}(X) \geq \mathtt{mp} \wedge P_{\geq \mathtt{ms}}(Y) \geq \mathtt{mp}$ を満たす，すなわち，X と Y が共に確率的頻出アイテム集合である必要がある．従って，上述の手順で確率的負の相関ルールの前件，後件を限定しても，抽出漏れが起きることはない．以下，それぞれの手順について，詳細を示す．

4.2.1　確率的頻出アイテム集合の発見

確率的頻出アイテム集合の発見に関しては，これまでに多くの手法が提案されている [11, 3, 12]．

本研究では，主に文献 [11] で提案された手法を用い，確率的頻出アイテム集合の抽出を行う．具体的には，接尾辞木 [7, 8, 9] として表現される集合列挙木を考え，その上を深さ優先で巡回する．その際，[11] や [12] に従い以下の 2 つの枝刈りを採用する．

第一の枝刈りは，最小確率に基づく枝刈りである．2 つのアイテム集合 X と X' に対し，$X \subset X'$ であるならば，それらの頻出確率は，$P_{\geq \mathtt{ms}}(X) \geq P_{\geq \mathtt{ms}}(X')$ を満たす．これにより，$P_{\geq \mathtt{ms}}(X) < \mathtt{mp}$，すなわちアイテム集合 X の確率が最小確率未満の場合，X' が確率的頻出アイテム集合になることはない．従って，接尾辞木上の探索により X から導出される X' に対し枝刈りを適用することができる．

第二の枝刈りは，チェルノフの不等式 [10] に基づく枝刈りである．$\mu = esup(X), \sigma = \frac{\mathtt{ms}-\mu-1}{\mu}$ としたとき，アイテム集合 X が

1. $\sigma \geq 2e-1 \quad \wedge \quad 2^{-\sigma\mu} < \mathtt{mp}$
2. $0 < \sigma < 2e-1 \quad \wedge \quad e^{\frac{-\sigma^2\mu}{4}} < \mathtt{mp}$

$$
\begin{aligned}
&P(X \Rightarrow \neg Y) \\
&= P\left[sup(X\neg Y) \geq \mathtt{ms} \wedge \tfrac{sup(X\neg Y)}{sup(X)} \geq \mathtt{mc} \wedge sup(XY) < \mathtt{ms} \wedge sup(X) \geq \mathtt{ms} \wedge sup(Y) \geq \mathtt{ms}\right] \\
&= P\left[sup(X\neg Y) \geq \mathtt{ms} \wedge \tfrac{sup(X\neg Y)}{sup(XY)+sup(X\neg Y)} \geq \mathtt{mc} \wedge sup(XY) < \mathtt{ms} \wedge sup(X) \geq \mathtt{ms} \wedge sup(Y) \geq \mathtt{ms}\right] \\
&= P\left[sup(X\neg Y) \geq \mathtt{ms} \wedge sup(XY) \leq \tfrac{1-\mathtt{mc}}{\mathtt{mc}} sup(X\neg Y) \wedge sup(XY) + 1 \leq \mathtt{ms} \wedge sup(X) \geq \mathtt{ms} \wedge sup(Y) \geq \mathtt{ms}\right] \\
&= P\left[sup(X\neg Y) \geq \mathtt{ms} \wedge sup(XY) \leq \tfrac{1-\mathtt{mc}}{\mathtt{mc}} sup(X\neg Y) \wedge sup(XY) + 1 \leq \mathtt{ms} \wedge sup(Y) \geq \mathtt{ms}\right] \\
&= P\left[sup(X\neg Y) \geq \mathtt{ms} \wedge sup(XY) \leq \tfrac{1-\mathtt{mc}}{\mathtt{mc}} sup(X\neg Y) \wedge sup(XY) \leq \mathtt{ms} - 1 \wedge sup(XY) + sup(\neg XY) \geq \mathtt{ms}\right] \\
&= P\left[sup(X\neg Y) \geq \mathtt{ms} \wedge sup(XY) \leq \min\left(\tfrac{1-\mathtt{mc}}{\mathtt{mc}} sup(X\neg Y), \mathtt{ms} - 1\right) \wedge sup(XY) + sup(\neg XY) \geq \mathtt{ms}\right] \\
&= \sum_{i=\mathtt{ms}}^{N} \sum_{j=0}^{\min\left(\mathtt{ms}-1, \frac{(1-\mathtt{mc})\cdot i}{\mathtt{mc}}\right)} \sum_{k=\mathtt{ms}-j}^{N} P\left[sup(X\neg Y) = i \wedge sup(XY) = j \wedge sup(\neg XY) = k\right] \\
&= \sum_{i=\mathtt{ms}}^{N} \sum_{j=0}^{\min\left(\mathtt{ms}-1, \frac{(1-\mathtt{mc})\cdot i}{\mathtt{mc}}\right)} \sum_{k=\mathtt{ms}-j}^{N} P\left[sup(X\neg Y) = i\right]\ P\left[sup(XY) = j\right]\ P\left[sup(\neg XY) = k\right] \\
&= \sum_{i=\mathtt{ms}}^{N} P_i(X\neg Y) \sum_{j=0}^{\min\left(\mathtt{ms}-1, \frac{(1-\mathtt{mc})\cdot i}{\mathtt{mc}}\right)} P_j(XY) \sum_{k=\mathtt{ms}-j}^{N} P_k(\neg XY)
\end{aligned}
$$

図 1: $P(X \Rightarrow \neg Y)$ の計算

のいずれかを満たす場合，$P_{\geq \mathtt{ms}}(X) < \mathtt{mp}$ となり，X は確率的頻出パターンとはならない．このことと頻出確率の単調性を利用し，X とそこから導出される X' に対し枝刈りを適用することができる．なお，上記の条件の確認は，実際の $P_{\geq \mathtt{ms}}(X)$ を計算する前に行うことが可能である．従って，まず上記の 2 条件を確認し，そこで枝刈りされなかったアイテム集合に対して実際の確率 $P_{\geq \mathtt{ms}}(X)$ を計算することになる．

一方，$P_{\geq \mathtt{ms}}(X)$ を求めるのに必要な $P_i(X)$ の計算に関しては，動的計画法を用いる方法や分割統治法に基づく方法が提案されている [11]．特に，分割統治法に基づく方法に関しては，フーリエ変換を用いた畳み込み積分を利用することで $O(N\log(N))$ 時間で全 $P_i(X)$ $(0 \leq i \leq N)$ を計算することが可能である．

抽出された各確率的頻出アイテム集合は，探索で用いられた接尾辞木上に配置され，次のステップである，確率的負の相関ルール抽出に利用される．

4.2.2 確率的負の相関ルールの発見

原理的には，接尾辞木中の確率的頻出アイテム集合の全組み合わせを検査することで，確率的負の相関ルールを抽出することができる．その際，考慮する組み合わせを如何に減らすかが重要となる．本研究では，抽出すべき確率的負の相関ルールに対し，極小性を要請している．既存研究 [7, 8, 9] でも示されているように，接尾辞木は，極小性の判定に関して有効なデータ構造である．子ノード間の優先順位を辞書順として接尾辞木を深さ優先に巡回した場合，アイテム集合 X が配置されたノードは，X の部分集合 $X' \subset X$ が配置されたノードをすべて巡回した後に巡回されることになる．この性質を，確率的負の相関ルール発見に応用する．

接尾辞木から確率的頻出アイテム集合 X を一つ選択し，それを確率的負の相関ルールの前件（もしくは後件）とする．なお，前件後件共に負のアイテム集合からなるルールを抽出する場合は，$\neg X$ を前件とする．そして，接尾辞木を根から深さ優先に巡回しながら，前件（もしくは後件）として X を持つルールを列挙していく．具体的には，接尾辞木の各ノードに配置された確率的頻出アイテム集合 Y を獲得し，ルール $X \Rightarrow \neg Y$ (もしくは $\neg Y \Rightarrow X$) を生成する．このとき，$X \Rightarrow \neg Y$ (もしくは $\neg Y \Rightarrow X$) が有効な負の相関ルールであれば，Y より先の部分木の巡回して得られる確率的頻出アイテム集合 Y' は Y の上位集合 $(Y \subset Y')$ となるので，そこから得られる確率的負の相関ルール $X \Rightarrow \neg Y'$ (もしくは $\neg Y' \Rightarrow X$) は極小性を満たすことはない．従って，枝刈りの対象とすることが可能である．

以上の考えに従い，$X \Rightarrow \neg Y$ の形式をした極小かつ有効な確率的負の相関ルールを導出するアルゴリズムの概要を図 2 に示す．なお図中では，簡略化のため，接尾辞木のノードとそこに含まれる確率的頻出アイテム集合を同一視している．また，$\neg X \Rightarrow Y$ 及び $\neg X \Rightarrow \neg Y$ の形式をしたルールに関しても，同様のアルゴリズムを構築することが可能である．

アルゴリズムでは，まず，すべての確率的頻出アイテム集合を獲得し，接尾辞木 $Tree$ 上に配置する．次に，結果を格納する変数 R を準備した後，接尾辞木を辿りながら前件となる確率的頻出アイテム集合 X を選択し，

入力：不確実データベース UDB，閾値 ms, mc, mp

1: すべての確率的頻出アイテム集合 X を含む接尾辞木 $Tree$ を構築する
2: $R := \emptyset$
3: $Tree$ を深さ優先に巡回しながら，各ノード X に対して，P_NAR(X, Tree, R) を実行する
4: R を出力する

P_NAR(X, Y, R)

1: **for each** Y' of Y's children
2: **if** $X \cap Y' = \emptyset \quad \wedge$
3: $\nexists X' \Rightarrow \neg Y'' \in R$ s.t. $X' \subseteq X \wedge Y'' \subseteq Y' \wedge \neg(X' = X \wedge Y'' = Y')$ **then**
3: **if** $P_{\geq \mathtt{ms}}(XY') < \mathtt{mp}$ **then**
4: $prob := P(X \Rightarrow \neg Y')$
5: **if** $prob \geq \mathtt{mp}$ **then**
6: $R := R \cup \{X \Rightarrow \neg Y'\}$
7: **continue**
8: **end if**
9: **end if**
10: $P\_NAR(X, Y', R)$
11: **end if**
12: **end for**

図 2: $X \Rightarrow \neg Y$ 形式の確率的負の相関ルール抽出アルゴリズム

手続き P_NAR(X, Tree, R) を呼びだすことで，確率的負の相関ルールを求める．手続き P_NAR(X, Tree, R) では，ノード Y の各子ノード Y' に対し，以下の処理を繰り返す．

- X との重複（$X \cap Y'$）を検査する．もし重複があれば，Y' の子孫 Y'' に対しても $X \cap Y'' \neq \emptyset$ が成り立つので，Y' 以降の探索をやめる．
- ルール $X \Rightarrow \neg Y'$ の極小性を検査する．既に $X \Rightarrow \neg Y'$ よりも極小なルールが発見されている場合は，以降の探索をやめる．
- アイテム集合 XY' が，確率的頻出アイテム集合でないこと（$P_{\geq \mathtt{ms}}(XY') < \mathtt{mp}$）を確認した上で，実際の確率 $P(X \Rightarrow \neg Y')$ を計算する．このとき，$P(X \Rightarrow \neg Y') \geq \mathtt{mp}$ ならば，$X \Rightarrow \neg Y$ を有効な確率的負の相関ルールとして R へと追加する．また，以降の探索をやめる．
- アイテム集合 XY' が確率的頻出アイテム集合である場合や，$X \Rightarrow Y$ が有効な確率的負の相関ルールでない場合は，再帰呼び出しを行い，Y' の子を対象とした探索を継続する．

ここで，$P(C_X \Rightarrow C_Y)$ の性質について簡単に考察する．確率的頻出アイテム集合 X と Y，Y' に対し，$Y \subset Y'$ ならば $P(X \Rightarrow \neg Y) \leq P(X \Rightarrow \neg Y')$ が成り立つ．従って，$P(X \Rightarrow \neg Y)$ が有効な確率的負の相関ルールであれば，$P(X \Rightarrow \neg Y')$ も有効な確率的負の相関ルールとなる．一方で，$P(\neg Y \Rightarrow X)$ と $P(\neg Y' \Rightarrow X)$ に関しては，単調性（$P(\neg Y \Rightarrow X) \leq P(\neg Y' \Rightarrow X)$）が成り立たない．従って，たとえ $\neg Y \Rightarrow X$ が有効な確率的負の相関ルールであっても，$\neg Y' \Rightarrow X$ が有効な確率的負の相関ルールとなるとは限らない．

本研究では，興味深さの判断基準として極小性を採用しているが，今後は，別の観点での特徴付けも必要になると考えている．

5 評価実験

Java 言語を用いて提案アルゴリズムを実装し，評価実験を行った．なお今回は，$X \Rightarrow \neg Y$ の形式をしたルールのみを対象とした．

実験では，Frequent Itemset Mining Dataset Repository [1] から入手した retail データ [4] から 500 件を取り出し，下記の手順で存在確率を付与した．

1. 平均 1，分散 0.1 のガウス分布に従い確率 p を求める．
2. p の値が 0 未満もしくは 2 を超える場合は，1. へ戻り，p を求めなおす．
3. p の値が 1 を超える場合は，$p = 2 - p$ とする．

実験に使用した PC のスペックは，OS：Windows 7, CPU：Xeon 3.40GHz，メインメモリ：32GB である．実験結果を表 1 に示す．表中において，FP は発見された確率的頻出アイテム集合の数，NAR は発見された極小な確率的負の相関ルールの数を表す．なお，すべての場合において，実行は数秒以内に終了した．

表から，最小確率 mp=0.80 の場合，最小支持度 ms の減少に伴い抽出される確率的負の相関ルールの数が増加していることが分かる．一方で mp=0.20 の場合は，ms の減少に伴い，抽出されるルール数が減少している．これは，以下の原因によるものと考えられる．最小確率 mp を固定した場合，最小支持度 ms の減少により，確率的負の相関ルールを構成するアイテム集合の数は増加する．一方で，頻出でない前件後件の組み合わせは減少してしまう．すなわち，ms の減少は，$P_{\geq \mathtt{ms}}(X) \geq \mathtt{mp}$

[1] http://fimi.ua.ac.be/

表 1: 実験結果

		mc =	.80	.60	.40	.20
mp	ms	FP	NAR			
.80	40	7	0	1	1	1
	30	10	1	2	2	2
	20	13	3	4	4	4
	10	29	3	4	4	4
.60	40	8	3	3	3	3
	30	10	4	4	4	4
	20	13	6	6	6	6
	10	31	6	6	6	6
.40	40	8	7	7	8	8
	30	10	7	7	8	8
	20	14	11	11	11	11
	10	31	11	11	11	11
.20	40	8	29	34	34	34
	30	10	29	35	35	35
	20	14	26	32	32	32
	10	31	26	32	32	32

を満たすアイテム集合 X の増加させる一方で，ルール $X \Rightarrow \neg Y$ が確率的負の相関ルールとなる条件の一つである $P_{\geq \mathtt{ms}}(XY) < \mathtt{mp}$ を満たす組み合わせを減少させる．このことにより，支持度の変化と得られるルール数に単調な関係が成り立たなくなっている．

6 まとめ

本研究では，既存の確率的頻出相関ルール発見手法 [11] と負の相関ルール発見手法 [7] を拡張し，不確実データベースから極小かつ有効な負の相関ルールを発見するアルゴリズムを提案した．また，$X \Rightarrow \neg Y$ の形式をしたルールを発見するシステムを実装し，初期的な評価実験を行った．

今後の課題としては，$\neg X \Rightarrow Y$ や $\neg X \Rightarrow \neg Y$ の形式をした負の相関ルール導出システムの開発及び種々のデータを利用しシステムの評価があげられる．

参考文献

[1] C. C. Aggarwal. *Managing and Mining Uncertain Data*. Springer, 2009.

[2] C. C. Aggarwal and P. S. Yu. A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623, 2009.

[3] T. Bernecker, H.-P. Kriegel, M. Renz,k F. Verhein and A. Zuefle : Probabilistic Frequent Itemset Mining in Uncertain Databases, *Proc.* of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.119-128 (2009)

[4] T. Brijs, G. Swinnen, K. Vanhoof and G. Wets: Using Association Rules for Product Assortment Decisions: A Case Study, *Proc.* of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp.254-260 (1999)

[5] C. Cornelis, P. Yan, X. Zhang and G. Chen : Mining Positive and Negative Association Rules from Large Databases, *Proc.* of the 2006 IEEE Conference on Cybernetics and Intelligent Systems, pp.613-618 (2006)

[6] N. Dalvi and D. Suciu. Efficient Query Evaluation on Probabilistic Databases. *The VLDB Journal*, 16(22):523–544 (2007)

[7] 井出典子，岩沼宏治，山本泰生 : 極小性を用いた負の相関ルールの効率的な抽出法，第 27 回人工知能学会全国大会 (2013)

[8] 亀谷由隆, 佐藤泰介 : 最小サポート上昇に基づく上位 k 関連パターン発見，第 1 回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 (2011)

[9] J. Li, H. Li, L. Wong, J. Pei and G. Dong : Minimum Description Length Principle: Generators are Preferable to Closed Patterns, *Proc.* of the 21st National Conference on Artificial Intelligence, Vol. 1, pp.409-414 (2006)

[10] R. Motwani and P. Raghavan : *Randomized Algorithms*, Cambridge University Press (1995)

[11] L. Sun, R. Change, D. W. Cheung and J. Cheng : Mining Uncertain Data with Probabilistic Guarantees, *Proc.* of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.273-282 (2010)

[12] Y. Tong, L. Chen and B. Ding : Discovering Threshold-based Frequent Closed Itemsets over Probabilistic Data, *Proc.* of the 2012 IEEE 28th International Conference on Data Engineering, pp.270-281 (2012)

[13] H. Wnag, X. Zhang and G. Chen : Mining a Complete set of Both Positive and Negative Association Rules from Large Databases, *Proc.* of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp.777-784 (2008)

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