サブパターンとスーパーパターンからの推定頻度に基づくパターンの興味深さの尺度の評価

吉田由起子† 太田 唯子† 小林 健一† 湯上 伸弘†

† (株) 富士通研究所 〒 211–8588 神奈川県川崎市中原区上小田中 4–1–1
E-mail: †{y-yoshida,yuiko,kenichi,yugami}@jp.fujitsu.com

あらまし データベースから発見される膨大なパターン群の中から, 興味深いパターン群を選び出すアプローチが注目されている. 著者たちは, パターンの出現頻度が, そのサブパターンとスーパーパターンの出現頻度から推定される出現頻度とどれくらい乖離しているかによって, パターンの興味深さを評価する sub+super 法を提唱している. ある種のデータベースにおいて, sub+super 法によって選択されたパターン群は, 従来手法よりも冗長性が非常に低く, データベース内を同等以上に広い範囲で被覆することができる. 本稿では, sub+super 法で選択されるパターン群の特徴と sub+super 法の適用に向いているデータベースの条件について議論する.

キーワード 興味深さの尺度, 頻出パターン, 相関性, アイテム集合マイニング, データマイニング

Evaluation of an interestingness measure of patterns based on frequencies estimated from their subpatterns and superpatterns

Yukiko YOSHIDA†, Yuiko OHTA†, Ken'ichi KOBAYASHI†, and Nobuhiro YUGAMI†

† Fujitsu Laboratories, Ltd Kamikodanaka 4–1–1, Nakahara-ku, Kawasaki, Kanagawa 211–8588 Japan
E-mail: †{y-yoshida,yuiko,kenichi,yugami}@jp.fujitsu.com

Abstract In knowledge discovery in databases, the number of discovered patterns is often too enormous for human to understand. Therefore, it is needed to select only useful, interesting patterns from them. For this purpose, we have proposed the *sub+super* method that measures the interestingness of a pattern based on how its actual frequency is higher than those estimated from its subpatterns and superpatterns. Compared to other existing methods, the *sub+super* method has ability to select such a group of interesting patterns that are much less redundant and cover as many features in the database. The ability is demonstrated strongly in certain types of databases. In this paper, we discuss the types of databases for which the *sub+super* method can be suited and the features of selected patterns in comparison to other existing methods.

Key words interestingness measures, frequent patterns, association, itemset mining, data mining

1. はじめに

データベースからの知識発見では, 一般的に, Apriori [1] 等の手法を用いて所定の最小サポート (minimum support) より多く出現するパターン (頻出パターン) を抽出することが行なわれているが, しばしば人間が把握しきれないほど膨大な数のパターンが抽出されてしまうことがある. 最小頻度の値を大きくすればパターンの抽出数を減らすことができるが, その代わりに, 頻度が低めの有用なパターンが見落とされがちになる. そこで, 膨大な数の頻出パターンの中から, 頻度とは別の基準で興味深いパターンを選び出すアプローチが注目されている [5].

本稿では, 最も基本的なパターンであるアイテム集合を対象として, その選択手法を取扱う. 既存手法としては, アイテム集合のサイズ別に出現頻度の上位から一定数ずつアイテム集合を選択する [4], アイテム集合についてその部分集合の相互依存度の高いものを選択する [6], アイテム集合の頻度がその部分集合の頻度情報だけで計算可能にならないものを選択する [3] などのアプローチが存在する. ところが, [4] のようにパターンのサイズ別に高頻度パターンの選択を行なうアプローチ, あるいは [3], [6] のように評価対象のパターンの頻度をその構成要素 (サブパターン) の頻度から推定するというアプローチの場合, 共起性が非常に高いアイテム群が存在するデータベースでは, それらのアイテムの組合せでできる多数のパターン群を同等に「興味深い」と解釈して選択してしまう傾向がある. しかし, そのようなパターン群の多くは冗長である.

この問題を解決する方法として, 著者たちは, 評価対象のパ

T_1 :	$\{A, B, C, D, E\}$
T_2 :	$\{D, E\}$
T_3 :	$\{E\}$
T_4 :	$\{A, B, C, D\}$
T_5 :	$\{D, E\}$
T_6 :	$\{A, B, C\}$

表 1　トランザクションのデータベース

$\{D\}$	4	$\{A, B\}$	3	$\{A, B, C\}$	3	$\{A, B, C, D\}$	2	$\{A, B, C, D, E\}$	1
$\{E\}$	4	$\{A, C\}$	3	$\{A, B, D\}$	2	$\{A, B, D, E\}$	1		
$\{A\}$	3	$\{B, C\}$	3	$\{A, C, D\}$	2	$\{A, B, C, E\}$	1		
$\{B\}$	3	$\{D, E\}$	3	$\{B, C, D\}$	2	$\{A, C, D, E\}$	1		
$\{C\}$	3	$\{A, D\}$	2	$\{A, B, E\}$	1	$\{B, C, D, E\}$	1		
		$\{B, D\}$	2	$\{A, C, E\}$	1				
		$\{C, D\}$	2	$\{A, D, E\}$	1				
		$\{A, E\}$	1	$\{B, C, E\}$	1				
		$\{B, E\}$	1	$\{B, D, E\}$	1				
		$\{C, E\}$	1	$\{C, D, E\}$	1				

表 2　パターンの頻度情報

ターンについて, そのサブパターンとスーパーパターン (評価対象のパターンを包含するパターン) から推定される頻度と実際の頻度がどれくらい乖離しているかによって, パターンの興味深さを評価する sub+super 法を提案した [8]. sub+super 法の特徴であるスーパーパターンからの頻度推定は, 共起性の高いアイテム群の組合せでできるパターン群の中で冗長なパターンに対しては興味深さを低く抑える作用があり, そのため, sub+super 法によって選択されたパターン群は冗長性が非常に低く, データベース内の様々な特徴を効率的に表現することができる.

本稿では, sub+super 法で選択されるパターン群の特徴と sub+super 法の適用に向いているデータベースの条件について議論する. 以下の構成は, 第 2 節で sub+super 法について説明し, 第 3 節では sub+super 法と他の既存手法による興味深いパターンの選択の実験および評価について, 第 4 節では sub+super 法が適するデータベースの条件, および sub+super 法の問題点について議論する. 第 5 節は本稿のまとめである.

2. sub+super 法

2.1 頻出アイテム集合マイニングの諸概念

ここでは, 以下の議論で用いる頻出アイテム集合マイニングの諸概念を説明する. $A = \{a_1, a_2, \cdots, a_Z\}$ をアイテム a_i の全体とする. $\mathcal{D}$ を, A の部分集合であるトランザクションで構成されたデータベースとする. A の部分集合をパターンと呼び, パターンを構成するアイテムの個数をパターンのサイズと呼ぶ. パターン $\boldsymbol{s}$ の頻度 $f(\boldsymbol{s})$ とは, データベース $\mathcal{D}$ 内でパターン $\boldsymbol{s}$ を含むトランザクションの数である. 所与の閾値 $minsup$ について, $f(\boldsymbol{s}) \geq minsup$ を満たすパターンを頻出パターンと呼び, $minsup$ を最小サポートと呼ぶ. パターン $\boldsymbol{s}$ がパターン $\boldsymbol{t}$ の真部分集合であるとき, $\boldsymbol{s}$ を $\boldsymbol{t}$ のサブパターンと呼び, $\boldsymbol{t}$ を $\boldsymbol{s}$ のスーパーパターンと呼ぶ.

2.2 パターンの興味深さの尺度

評価対象のパターン $\boldsymbol{s}$ に対して, $\boldsymbol{s}^-$ を空でないサブパターン, $\boldsymbol{s}^+$ をスーパーパターンとする. $\boldsymbol{s}$ に属し $\boldsymbol{s}^-$ に属さないアイテムの集合で構成されるパターンを $\boldsymbol{s} \setminus \boldsymbol{s}^-$ で表し, $\boldsymbol{s}^+$ に属し $\boldsymbol{s}$ に属さないアイテムの集合で構成されるパターンを $\boldsymbol{s}^+ \setminus \boldsymbol{s}$ で表す. パターン $\boldsymbol{s}^-$ および $\boldsymbol{s} \setminus \boldsymbol{s}^-$ の出現頻度の独立性を仮定すると, $\boldsymbol{s}^-$ および $\boldsymbol{s} \setminus \boldsymbol{s}^-$ の頻度情報が与えられたときの $\boldsymbol{s}$ の推定頻度 $\hat{f}(\boldsymbol{s}|\boldsymbol{s}^-)$ を次式で計算することができる:

$$\hat{f}(\boldsymbol{s}|\boldsymbol{s}^-) = f(\boldsymbol{s}^-) \cdot \frac{f(\boldsymbol{s} \setminus \boldsymbol{s}^-)}{N_{\mathcal{D}}}$$

ここで $N_{\mathcal{D}}$ はデータベース $\mathcal{D}$ 内のトランザクション数である. また, $\boldsymbol{s}$ および $\boldsymbol{s}^+ \setminus \boldsymbol{s}$ の出現頻度の独立性を仮定すると, $\boldsymbol{s}^+$ および $\boldsymbol{s}^+ \setminus \boldsymbol{s}$ の頻度情報が与えられたときの $\boldsymbol{s}$ の推定頻度 $\hat{f}(\boldsymbol{s}|\boldsymbol{s}^+)$ を次式で計算することができる:

$$\hat{f}(\boldsymbol{s}|\boldsymbol{s}^+) = f(\boldsymbol{s}^+) \cdot \frac{N_{\mathcal{D}}}{f(\boldsymbol{s}^+ \setminus \boldsymbol{s})}$$

既存の興味深いパターンの選択手法の多くは, 基本的にはサブパターンの頻度からパターンの推定頻度を計算し, パターンの実際の頻度と推定頻度との差異が大きければ興味深いパターンとみなすというものである. ところが, そのような方法の場合, 共起性が非常に高いアイテム群が存在するデータベースでは, それらのアイテム群の組合せ (それらのアイテム群全体で構成されるパターンのサブパターン群) の多くを「興味深い」と解釈してしまう傾向がある. そのようなパターン群の多くは冗長である.

2.3 sub+super 法の興味深さの尺度

この問題を回避するための手法として, 頻出パターンの中から, 次式で表されるパターンの興味深さの尺度 $I_{sub+super}$ に基づいて興味深いパターンを選択する sub+super 法を導入する:

$$I_{sub+super}(\boldsymbol{s}) = \frac{1}{\pi}\left[\min\left\{\arctan\left(\frac{f(\boldsymbol{s})}{\hat{f}(\boldsymbol{s}|\boldsymbol{s}^-)}\right) : \boldsymbol{s}^- \in S^-\right\} + \min\left\{\arctan\left(\frac{f(\boldsymbol{s})}{\hat{f}(\boldsymbol{s}|\boldsymbol{s}^+)}\right) : \boldsymbol{s}^+ \in S^+\right\}\right]$$

ここで, S^- および S^+ はパターン $\boldsymbol{s}$ に対して所定の方法で生成されるサブパターンおよびスーパーパターンの集合である. この尺度では, サブパターンだけではなくスーパーパターンと組み合わせてパターンの推定頻度を計算し, パターンの実際の頻度が推定頻度よりも大きいほど興味深いと評価する. また, パターンの実際の頻度が, どのサブパターン, スーパーパターンからの推定頻度と比べてもじゅうぶん大きくなっている場合にのみ興味深いと評価するために, サブパターン群とスーパーパターン群における実際の頻度と推定頻度との比の最小値で評価し, さらに, 実際の頻度と推定頻度との比を有限値に抑えるために, arctan で評価している. $1/\pi$ は正規化のための係数である.

評価対象のパターン $\boldsymbol{s}$ に対するサブパターン $\boldsymbol{s}^-$ および $\boldsymbol{s}^+$ については種々の生成方法が考えられるが, 以下の説明では, $\boldsymbol{s}$ よりもパターンのサイズが 1 だけ小さいサブパターン群と, 1 だけ大きいスーパーパターン群を用いることとする.

表 1 に示す簡単なデータベースを用いて, sub+super 法の計算例を示す. このデータベース内のパターンの頻度を数え上げると, 表 2 に示すパターンの頻度情報が得られる. パターン $\{A,B,C\}$ を評価対象とすると, サブパターン $\{B,C\}$, $\{A,C\}$, $\{A,B\}$ の頻度情報からそれぞれ, パターン $\{A,B,C\}$ の推定頻度が

$$\hat{f}(\{A,B,C\}|\{B,C\}) = f(\{B,C\}) \cdot \frac{f(\{A\})}{N_{\mathcal{D}}} = 3 \cdot \frac{3}{6} = 1.5$$
$$\hat{f}(\{A,B,C\}|\{A,C\}) = f(\{A,C\}) \cdot \frac{f(\{B\})}{N_{\mathcal{D}}} = 3 \cdot \frac{3}{6} = 1.5$$
$$\hat{f}(\{A,B,C\}|\{A,B\}) = f(\{A,B\}) \cdot \frac{f(\{C\})}{N_{\mathcal{D}}} = 3 \cdot \frac{3}{6} = 1.5$$

のように計算される. スーパーパターン $\{A,B,C,D\}$, $\{A,B,C,E\}$ の頻度情報からそれぞれ, パターン $\{A,B,C\}$ の推定頻度が

$$\hat{f}(\{A,B,C\}|\{A,B,C,D\}) = f(\{A,B,C,D\}) \cdot \frac{N}{f(\{D\})} = 2 \cdot \frac{6}{4} = 3$$
$$\hat{f}(\{A,B,C\}|\{A,B,C,E\}) = f(\{A,B,C,E\}) \cdot \frac{N}{f(\{E\})} = 1 \cdot \frac{6}{4} = 1.5$$

のように計算される.

サブパターン $\{B,C\}$, $\{A,C\}$, $\{A,B\}$, およびスーパーパターン $\{A,B,C,D\}$, $\{A,B,C,E\}$ からのパターン $\{A,B,C\}$ の推定頻度に対する実際の頻度 $f(\{A,B,C\})$ の比の arctan はそれぞれ,

$$\arctan\left(\frac{f(\{A,B,C\})}{\hat{f}(\{A,B,C\}|\{B,C\})}\right) = \arctan\left(\frac{3}{1.5}\right) = 1.107$$
$$\arctan\left(\frac{f(\{A,B,C\})}{\hat{f}(\{A,B,C\}|\{A,C\})}\right) = \arctan\left(\frac{3}{1.5}\right) = 1.107$$
$$\arctan\left(\frac{f(\{A,B,C\})}{\hat{f}(\{A,B,C\}|\{A,B\})}\right) = \arctan\left(\frac{3}{1.5}\right) = 1.107$$
$$\arctan\left(\frac{f(\{A,B,C\})}{\hat{f}(\{A,B,C\}|\{A,B,C,D\})}\right) = \arctan\left(\frac{3}{3}\right) = 0.785$$
$$\arctan\left(\frac{f(\{A,B,C\})}{\hat{f}(\{A,B,C\}|\{A,B,C,E\})}\right) = \arctan\left(\frac{3}{1.5}\right) = 1.107$$

と計算される. したがって, パターン $\{A,B,C\}$ の興味深さの尺度は

$$I_{sub+super}(\{A,B,C\}) = \frac{1}{\pi}(1.107 + 0.785) = 0.602$$

となる.

サブパターンだけで比較する尺度では, $\{B,C\}$, $\{A,C\}$, $\{A,B\}$ には, $\{A,B,C\}$ と同等の評価値が与えられる. これは, $\{D,E\}$ や $\{A,B,C,D\}$ よりも評価値が高い. しかし, もとのデータベース (表 1) を見れば分かるように, アイテム A, B, C はつねに共起しているので, 本来, パターン $\{A,B,C\}$ を $\{B,C\}$, $\{A,C\}$, $\{A,B\}$ よりも高く評価すべきであって, $\{A,B,C\}$, $\{B,C\}$, $\{A,C\}$, $\{A,B\}$ に同等の評価値を与えるのは適切ではない. しかも, サブパターンからの頻度推定のみでは, 別種のパターン $\{D,E\}$, $\{A,B,C,D\}$ の評価が相対的に低くなってしまい, $\{B,C\}$, $\{A,C\}$, $\{A,B\}$ に比べて選択されにくくなる.

それに対して, sub+super 法は, スーパーパターンからの頻度推定によって冗長なパターン $\{B,C\}$, $\{A,C\}$, $\{A,B\}$ の評価値を低く抑えることができる.

3. 実　　験

3.1 比較対象の手法

sub+super 法と他の手法を比較評価するために, 各手法がデータベースから興味深いパターンとしてどのようなアイテム集合を選択するか実験を行なった.

sub+super 法との比較対象の手法として下記の sub, N-most, および m-patterns 法を用い, 表 4 のデータベースに対して, 最小サポート $minsup$ を適宜設定して生成された頻出アイテム集合の中から, 各手法による興味深いパターンの上位 $M = 200$ 個を選択した:

- sub: sub+super 法の特徴であるスーパーパターンからの頻度推定の効果を調べるために, $I_{sub+super}$ 式から, スーパーパターンに関する部分を取り除いた尺度を作成した:

$$I_{sub}(\boldsymbol{s}) = \frac{2}{\pi} \min\left\{\arctan\left(\frac{f(\boldsymbol{s})}{\hat{f}(\boldsymbol{s}|\boldsymbol{s}^-)}\right) : \boldsymbol{s}^- \in S^-\right\}$$

- N-most (N-most interesting itemsets) [4]: アイテム集合のサイズ ($k = k_{min}, \cdots, k_{max}$) 別に, 出現頻度の高い順に N 個のアイテム集合を選択する手法. N-most によって選択されるアイテム集合の数は $(k_{max} - k_{min} + 1) \times N$ である.
- m-patterns (mutually dependent, frequent patterns) [6]: 頻出アイテム集合について, その部分集合の相互依存度が高いものを選択する手法. 相互依存度は, アイテム集合に属すアイテムの頻度に対する条件付確率によって定義される. 具体的には, $f(\boldsymbol{s}) \geqq minsup$ を満たす頻出アイテム集合の中から, 相互依存度 $\min\{f(\boldsymbol{s})/f(\{a\}) : a \in \boldsymbol{s}\}$ の高い順に M 個のアイテム集合を選択する.

3.2 実験用データベース

実験に用いたデータベースは, StatLog [7] (dna) および UCI Machine Learning Repository [2] (internet-ads, mushroom, soybean, zoo) から入手した機械学習用データベースを, アイテム集合のトランザクションとして取扱えるように変換したものである. データベース dna のもとのデータは, 180 個のブーレアン値属性および 3 値のクラス (1, 2, 3) のリストとして表現されてい

属性名 (英) (日)	animal name 動物名	hair 毛	feathers 羽	eggs 卵生	milk 哺乳性	airborne 飛行性	aquatic 水生	predator 捕食性	toothed 歯	backbone 背骨	breathes 肺呼吸	··· ···
トランザクション (変換前)	bear	1	0	0	1	0	0	1	1	1	1	···
	carp	0	0	1	0	0	1	0	1	1	0	···
	chicken	0	1	1	0	1	0	0	0	1	1	···

⇓

トランザクション (変換後)	毛=○:羽=×:卵生=×:哺乳性=○:飛行性=×:水生=×:捕食性=○:歯=○:背骨=○:肺呼吸=○:···
	毛=×:羽=×:卵生=○:哺乳性=×:飛行性=×:水生=○:捕食性=×:歯=○:背骨=○:肺呼吸=×:···
	毛=×:羽=○:卵生=○:哺乳性=×:飛行性=○:水生=×:捕食性=×:歯=×:背骨=○:肺呼吸=○:···

表 3　データベースのトランザクションの変換例

データベース	internet-ads	dna	mushroom	zoo	soybean	soybean2
トランザクション数	3279	2000	8124	101	307	307
トランザクションの (平均) サイズ	13.7 (平均)	46.6 (平均)	23 (一定)	17 (一定)	36 (一定)	16 (一定)
アイテムの種類	1557	183	119	43	117	35
アイテムの頻度の平均	0.00884	0.254	0.193	0.395	0.307	0.457
アイテムの頻度の分散	0.00105	0.00301	0.0559	0.0716	0.0904	0.171
最小サポート	0.025	0.05	0.05	0.05	0.3	0.05
パターンの最大サイズ	8	5	8	8	8	8
評価対象頻出パターン数	11,144	20,630	1,740,884	334,531	1,691,042	126,605

表 4　実験に用いたデータベース, および頻出パターン生成情報

データベース	internet-ads	dna	mushroom	zoo	soybean	soybean2
sub+super	0.212	0.296	0.652	0.813	0.703	0.539
sub のみ	0.124	0.153	0.084	0.306	0.173	0.348
N-most	0.283	0.295	0.634	0.844	0.565	0.933
m-patterns	0.258	0.343	0.656	0.881	0.435	0.800

表 5　選択されたパターン群によるデータベース内の被覆率 (Cover)

データベース	internet-ads	dna	mushroom	zoo	soybean	soybean2
sub+super	0.187	0.055	0.052	0.074	0.271	0.140
sub のみ	0.219	0.140	0.290	0.170	0.759	0.258
N-most	0.580	0.167	0.477	0.547	0.810	0.766
m-patterns	0.296	0.100	0.346	0.421	0.815	0.824

表 6　選択されたパターン群によるデータベース内の平均冗長率 (Redundancy)

る. データベース internet-ads のもとのデータは, 3 種類の連続値属性, 457 種類のブーレアン値属性, および 2 値のクラス (ad., nonad.) のリストとして表現されている. このリストからそれぞれ, ブーレアン値=1 となっている属性の名前およびクラス値を取り出したものを, アイテム集合のトランザクション・データとして利用した. データベース zoo のもとのデータは, 表 3 の上側に示すように, 18 種類のブーリアン値あるいは離散値属性のリストとして表現されている. 各リストの第 1 属性 (動物名) はリストの ID である. これを表 3 下側のように, 各属性と値との対応を「属性名=値」という形式のアイテムとして表現し, アイテム集合のトランザクションとして変換して利用した (「:」はトランザクション上のアイテムの区切りを表す). データベース mushroom, soybean についても zoo と同様に変換した.

さらに, soybean に基づいて別のデータベース soybean2 を作成した. soybean は, 19 値のクラスおよび 35 種類の離散値属性によるトランザクション 307 個で構成されているが, 頻度 80% 以上の属性値 (つまり, 307 個のトランザクションのうちの 246 個以上に共通に出現する属性値) が存在する属性のみを使用するようにトランザクションを再構成したものが soybean2 である. たとえば, soybean の 25 番目の属性 mycelium (菌糸体) の値の分布は, absent が 99.3% (305 個のトランザクション), present が 0.7% (2 個のトランザクション) となっていて, 頻度 80% 以上の属性値が存在するので, この属性を使用する. また, soybean の 2 番目の属性 plant-stand (植物の背丈) の値の分布は, normal が 54.7% (168 個のトランザクション), lt-normal が 45.3% (139 個のトランザクション) となっていて, 頻度 80% 以上の属性値が存在しないので, この属性を使用しない. 35 種類の離散値属性のうち上記の条件を満たしている属性は 16 種類あり, これらの属性は, soybean2 内の大部分 (80% 以上) のトランザクションで共通の値を取ることになる.

変換後のデータベース mushroom, zoo, soybean, および soybean2 は一定サイズのトランザクションで構成され, dna および internet-ads はいろいろなサイズのトランザクションで構成されている.

3.3 評価方法

各手法によって選択されたパターン群 $\mathcal{R} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_M\}$ を下記の 2 つの観点で評価した:

- データベース内の被覆率: 選択されたパターン群によって, データベース内の特徴をどれくらい多く表現することができるか. これを評価するための尺度として次式を用いた. これは, データベース内のトランザクション上で選択されたパターン群のどれかとマッチする箇所をアイテム数ベースでカウントし, データベース内のトランザクションのアイテム数の総和との比を取ったものである:

$$Cover = \frac{\sum_{t \in \mathcal{D}} k(\cup_{s \in \mathcal{R}, s \subseteqq t} s)}{\sum_{t \in \mathcal{D}} k(t)}$$

ここで, $k(s)$ はパターンあるいはトランザクション s に属すアイテムの数である.

- データベース内の平均冗長率: 選択されたパターン群によって, データベース内のトランザクションをどれくらい効率的に分類することができるか. たとえば, 2 つの異なるパターンが, データベース内の似たようなトランザクション群とマッチする場合と, まったく別々のトランザクション群とマッチする場合とでは, 後者のほうがより多くのトランザクションに関する情報が得られるので有用であると考えられる. これを評価するための尺度として次式を用いた.

$$Redundancy = \frac{2}{M(M-1)} \sum_{1 \leqq i < j \leqq M} \frac{f(\{\boldsymbol{t} \in \mathcal{D} : s_i \subseteqq t \text{ かつ } s_j \subseteqq t\})}{f(\{\boldsymbol{t} \in \mathcal{D} : s_i \subseteqq t \text{ あるいは } s_j \subseteqq t\})}$$

これは, 選択されたパターン群の各パターンのペアについて, 両者のどちらかがマッチするトランザクション群に対して両者ともにマッチするトランザクション群の割合をパターンのペアの総数で平均化したものである.

3.4 選択されたパターン群の特徴

表 5 は選択されたパターン群によるデータベース内の被覆率である. sub+super 法を他の手法と比較すると, データベース soybean では他の手法よりも被覆率がかなり高く, データベース mushroom, zoo, internet-ads では N-most および m-patterns と同程度かやや低く, データベース soybean2 では N-most および m-patterns と比べてかなり低いという結果になった. sub+super 法からスーパーパターンによる頻度推定部分を取り除いた sub 法は, 実験に用いた各データベースについて他の手法よりも被覆率がかなり低かった.

sub+super 法は, 共起性が非常に高いアイテム群が組み合さってできているパターン群の中では, サイズができるだけ大きいパターンを優先的に選択するような仕組みを持っている. 一般に, サイズが大きいパターンは, 小さいパターンに比べてデータベース内のトランザクションとマッチしにくいので, sub+super 法で選択されたパターン群はデータベース内での被覆率が低くなると思われたが, それにもかかわらず, sub+super 法で選択されたパターン群は, データベース soybean2 以外では他の手法と同程度かそれ以上の被覆率を示した.

データベース soybean2 では, 全 16 種類の属性が大部分 (80% 以上) のトランザクションで共通の値を取るという性質を持っているため, このデータベースについては, 単に出現頻度の高い順にパターンを選択するだけの N-most 法で, データベース内のほとんどすべての領域を被覆するパターン群を選び出すことができる. 一方, sub+super 法は, 単に頻出ではなく, 多様なパターン群を選択しようとするために, このようなデータベースでは低頻度のパターン群を選択することになり, N-most 法や (部分集合の相互依存度が高いパターンを選択する) m-patterns 法と比べると被覆率が低くなってしまう.

表 6 は選択されたパターン群によるデータベース内での平均冗長率である. 他の手法に比べて, sub+super 法は冗長率が非常に低いことが分かる. sub 法は, N-most および m-patterns よりも冗長率が低く, sub+super 法よりも冗長率が高かった. とくに, データベース soybean2 について比べると, sub+super 法は, N-most 法や m-patterns 法よりも被覆率は低いが, 平均冗長率はより優れている (冗長性が低いパターン群を選択することができる) ことが分かる.

選択されたパターン群の冗長性について具体例を挙げて議論する. 表 7–10 は, データベース zoo(注1) について, 各手法によって選択された興味深いパターン群の上位 16 個をリストアップし, 各パターンのペア (興味深さの上位 i, j 番目) の冗長率 (i, j のどちらかがマッチするトランザクション数に対する両者が重複してマッチするトランザクション数の割合) を表したものである. sub+super 法によって選択されたパターン群は, 他の手法に比べてサイズが大きく, 互いに非常に異なったアイテム群で構成されている. また, 各パターンのペアの冗長率が非常に低いことから, 1 つ 1 つのパターンがそれぞれ異なるトランザクション群を表していることが分かる. 一方, N-most 法は, パターンのサイズ k_{min} から k_{max} 別に頻度の上位のパターンを順番に選択する仕組みなので, 上位 16 個の中でも, 特定の頻出パターンのサブパターン群が繰り返し出現していることが分かる. N-most のパターン群の冗長率はきわめて高く, 興味深さの上位にランクされたいくつものパターン群が, 実は, 特定の互いに似通ったトランザクション群とマッチしているにすぎない. m-patterns 法で選択されたパターン群については, 共起性の高いアイテム群の組合せである 2 つのグループにまとめられる: 第 1 グループは 1, 4–6, 9–13, 15, 16 番目のパターンで, 哺乳類動物が共通に持つ特徴 (哺乳性である, 卵生ではない, 毛が生えている, 羽がない, 飛行性がない, 肺呼吸する, 背骨がある, 尾がある) の組合せであり, 第 グループは 2, 3, 7, 8, 12, 14, 15 番目のパターンで, 鳥が共通に持つ特徴 (哺乳性でない, 卵生である, 毛が生えていない, 羽がある, 肺呼吸する, 背骨がある, 尾がある) の組合せである. sub 法によって選択されたパターン群も同様に, 昆虫が共通に持つ特徴で構成される第 1 グループと, 魚が共通に持つ特徴で構成される第 2 グループのヴァリエーションにすぎない.

(注1): zoo は, クマ, コイ, ニワトリ等の 101 の動物の特徴を 16 個の属性で記述したデータであり, 17 番目の属性 (タイプ) にはその動物が属す種類 (哺乳類, 鳥, 魚, 昆虫等の 7 通り) の情報が与えられている.

これらの手法と比較すると, sub+super 法は, 共起性の非常に高いアイテム群の組合せでできる多数のパターン群の中から互いに似通った (冗長な) パターン群を排除することができ, また, 単に頻度の高い順にパターン群を選ぶ場合とは異なり, データベース内のトランザクションが持つ様々な特徴を効果的に表現できるパターン群を優先的に選択している. また, sub+super 法からスーパーパターンによる頻度推定項を削除した sub 法では, 被覆率が他の手法に比べてかなり低く, 冗長率が sub+super 法と N-most, m-patterns との中間くらいであることから, sub+super 法の性能には, スーパーパターンからの頻度推定が強く影響していることが分かる.

4. 議　　論

sub+super 法の枠組みと上記の実験・評価からの知見によって, sub+super 法に適したデータベースとして下記のような条件が考えられる:

- 共起性の高いアイテム群が存在すること: 従来手法では, そのようなアイテム群の組合せでできる多数のパターン群を同等に「興味深い」と解釈して選択してしまうが, sub+super 法は冗長性の低いパターン群のみ選択できる.
- ただし, 共起性の高いアイテム群は, 大部分のトランザクションに共通に出現するといった一様な分布ではなく, あるアイテム群は特定のトランザクション群で出現し, 別のアイテム群は別のトランザクション群で出現するといった分布の非一様性があること: N-most 法等と比べて, sub+super 法は別々のトランザクション群を特徴付ける別々のアイテム群を選択する能力が高い.
- また, トランザクションのサイズ (トランザクション内のアイテム数) が十分に大きいこと: sub+super 法はサイズができるだけ大きいパターンを優先的に選択するようになっているので, データベース内のトランザクションのサイズが小さいものばかりでは十分な効果を発揮することができないからである.

最後に, sub+super 法の計算の効率性について述べる. 頻出パターンの抽出手法として有名な Apriori 法では, あるパターンが頻出でなければ, そのパターンを含むスーパーパターンはすべて頻出でないという (パターン頻度の) downward closed の性質を利用して, 頻出パターンの候補の絞込みを効率的に行なっている. 既存の興味深さの尺度の中にも, downward closed の性質を満たし, Apriori と類似のアプローチで興味深いパターンの候補を効率的に絞り込むことができる手法が多い. ところが sub+super 法の興味深さの尺度 $I_{sub+super}$ は downward closed の性質を満たしていない (興味深くないパターンを含むスーパーパターンが興味深い場合がある) ので, Apriori 法のアプローチをそのまま適用して興味深いパターンの候補を絞り込むことはできない. また, sub+super 法では, パターンの興味深さを評価するためにはサブパターンだけではなくスーパーパターンの頻度情報を知る必要があり, それだけ多くの計算コストが要求される. このように, sub+super 法はパターンの評価の効率性が高くないので, 現在のところ大規模データベースへの適用には向いていない.

5. おわりに

本稿では, パターンの興味深さの評価方法として, サブパターンとスーパーパターンからの頻度推定を用いる sub+super 法を紹介し, 他の既存手法との比較実験を行い, sub+super 法の適用に向いているデータベースの性質および選択されたパターン群の特徴について議論した. sub+super 法は, トランザクション・サイズ (トランザクション内のアイテム数) が十分に大きく, 共起性の高いアイテム群が存在し, さらにそれらのアイテム群が非一様に分布するデータベースに適していて, sub+super 法によって選択されたパターン群は, 従来手法よりも冗長性が非常に低く, データベース内を同等以上に広い範囲で被覆することができることを示した.

現在の問題点として, sub+super 法の評価尺度は downward closed の性質を持たないので Apriori 法のようにはパターン候補の絞り込みができず, また, サブパターンだけではなくスーパーパターンからの評価も行う必要があるので, パターンの評価に非常に計算がかかる点が挙げられる.

今後の課題として, パターン評価を効率化できるように sub+super 法を改良し, アイテムの種類やトランザクションのサイズが非常に大きいデータベースに適用して有効性を評価すること, また, アイテム集合だけではなく, 相関ルールや順序列パターンの興味深さの評価に適用することを検討している.

文　　献

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th Int'l Conference on Very Large Databases (VLDB)*, 1994.

[2] C. Blake and C. Merz. UCI repository of machine learning databases, 1998. University of California, Irvine, Dept. of Information and Computer Sciences [http://www.ics.uci.edu/~mlearn/MLRepository.html].

[3] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *Proc. of the 13th European Conference on Machine Learning / the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, 2002.

[4] A. W.-C. Fu, R. W.-W. Kwong, and J. Tang. Mining n-most interesting itemsets. In *Proc. of the 12th Int'l Symposium on Methodologies for Intelligent Systems (ISMIS)*, 2000.

[5] R. J. Hilderman and H. J. Hamilton. *Knowledge Discovery and Measures of Interest*. Kluwer Academic Publishers, 2001.

[6] S. Ma and J. L. Hellerstein. Mining mutually dependent patterns. In *Proc. of the 2001 IEEE Int'l Conference on Data Mining (ICDM)*, 2001.

[7] D. Michie, D. Spiegelhalter, and C. Taylor. The StatLog datasets, 1994. Esprit Project 5170 StatLog (1991-94) [http://www.ncc.up.pt/liacc/ML/statlog/].

[8] Y. Yoshida, Y. Ohta, K. Kobayashi, and N. Yugami. Mining interesting patterns using estimated frequencies from subpatterns and superpatterns. In *Proc. of the 14th Int'l Conference on Algorithmic Learning Theory and the 6th Int'l Conference on Discovery Science (ALT/DS)*, 2003.

	パターンのペア (i, j) の冗長率															
パターン	i \ j	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
羽=×:卵生=○:哺乳性=×:毒=○	1	0.20	0.09	0.50	0	0	0	0.09	0.09	0.20	0.50	0	0	0	0	0
毛=○:羽=×:飛行性=○:水生=×:捕食性=×:肺呼吸=○:ひれ=×:ネコより大きい=×	2	-	0	0	0	0	0	0.09	0	0.50	0.09	0	0	0	0	0
羽=×:卵生=○:飛行性=×:水生=○:ひれ=×:足=4:家庭向き=×	3	-	-	0.09	0	0	0	0	0.09	0	0.09	0.09	0	0.09	0.09	0
毛=×:羽=×:哺乳性=×:飛行性=×:捕食性=○:毒=○:家庭向き=×	4	-	-	-	0	0	0	0	0.09	0	0.71	0	0	0	0	0
水生=×:捕食性=×:背骨=○:肺呼吸=○:毒=×:ひれ=×:尾=○:家庭向き=○	5	-	-	-	-	0.56	0.40	0.40	0	0	0	0	0	0	0	0
羽=×:飛行性=×:捕食性=×:歯=○:背骨=○:毒=×:尾=○:家庭向き=○	6	-	-	-	-	-	0.20	0.09	0	0	0	0	0	0	0	0
捕食性=×:背骨=○:毒=×:家庭向き=○:ネコより大きい=×	7	-	-	-	-	-	-	0.71	0	0.09	0	0	0	0	0.09	0.09
水生=×:捕食性=×:肺呼吸=○:ひれ=×:家庭向き=○:ネコより大きい=×	8	-	-	-	-	-	-	-	0	0.20	0	0	0	0	0.09	0.09
卵生=○:飛行性=×:水生=○:捕食性=○:背骨=○:尾=○:家庭向き=×:ネコより大きい=○	9	-	-	-	-	-	-	-	-	0	0	0	0	0.20	0	0
毛=○:羽=×:水生=×:捕食性=×:肺呼吸=○:ひれ=×:尾=×	10	-	-	-	-	-	-	-	-	-	0.09	0.09	0.09	0	0.09	0.20
羽=×:哺乳性=×:毒=○:ひれ=×:家庭向き=×:ネコより大きい=×	11	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0
羽=×:飛行性=×:捕食性=×:肺呼吸=○:毒=×:ひれ=×:尾=×:家庭向き=×	12	-	-	-	-	-	-	-	-	-	-	-	0.09	0	0.71	0.71
水生=×:背骨=○:肺呼吸=○:毒=×:ひれ=×:足=2:家庭向き=×:ネコより大きい=○	13	-	-	-	-	-	-	-	-	-	-	-	-	0.20	0	0.09
卵生=○:飛行性=×:歯=×:毒=×:ひれ=×:家庭向き=×:ネコより大きい=○	14	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
羽=×:飛行性=×:捕食性=×:肺呼吸=○:毒=×:ひれ=×:尾=×:ネコより大きい=×	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.71
羽=×:飛行性=×:水生=×:捕食性=×:肺呼吸=○:毒=×:ひれ=×:尾=×	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 7　sub+super 法によって選択されたパターン群, および各パターンのペアの冗長率

	パターンのペア (i, j) の冗長率															
パターン	i \ j	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
足=6:タイプ=昆虫	1	0	0	1	1	0.80	0	0	0	0	0	0	0	0	0	1
ひれ=○:タイプ=魚	2	-	0	0	0	0	0	1	1	1	1	1	1	0.81	1	0
背骨=×:タイプ=その他	3	-	-	0	0	0.11	0	0	0	0	0	0	0	0	0	0
背骨=×:タイプ=昆虫	4	-	-	-	1	0.80	0	0	0	0	0	0	0	0	0	1
背骨=×:足=6:タイプ=昆虫	5	-	-	-	-	0.80	0	0	0	0	0	0	0	0	0	1
背骨=×:足=6	6	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0.80
羽=○:タイプ=鳥	7	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0
肺呼吸=×:タイプ=魚	8	-	-	-	-	-	-	-	1	1	1	1	1	0.81	1	0
肺呼吸=×:ひれ=○:タイプ=魚	9	-	-	-	-	-	-	-	-	1	1	1	1	0.81	1	0
足=0:タイプ=魚	10	-	-	-	-	-	-	-	-	-	1	1	1	0.81	1	0
ひれ=○:足=0:タイプ=魚	11	-	-	-	-	-	-	-	-	-	-	1	1	0.81	1	0
肺呼吸=×:足=0:タイプ=魚	12	-	-	-	-	-	-	-	-	-	-	-	1	0.81	1	0
肺呼吸=×:ひれ=○:足=0:タイプ=魚	13	-	-	-	-	-	-	-	-	-	-	-	-	0.81	1	0
ひれ=○:足=0	14	-	-	-	-	-	-	-	-	-	-	-	-	-	0.81	0
肺呼吸=×:ひれ=○:足=0	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
尾=×:タイプ=昆虫	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 8　sub 法によって選択されたパターン群, および各パターンのペアの冗長率

	パターンのペア (i, j) の冗長率															
パターン	i \ j	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
毒=×	1	0.87	0.76	0.68	0.60	0.51	0.43	0.43	0.81	0.85	0.76	0.66	0.59	0.51	0.43	0.41
毒=×:家庭向き=×	2	-	0.65	0.57	0.52	0.58	0.36	0.36	0.92	0.72	0.67	0.75	0.51	0.42	0.49	0.34
肺呼吸=○:毒=×:ひれ=×	3	-	-	0.89	0.79	0.66	0.48	0.48	0.61	0.72	0.65	0.55	0.45	0.66	0.31	0.45
背骨=○:肺呼吸=○:毒=×:ひれ=×	4	-	-	-	0.89	0.75	0.54	0.54	0.53	0.80	0.72	0.61	0.49	0.75	0.34	0.51
背骨=○:肺呼吸=○:毒=×:ひれ=×:尾=○	5	-	-	-	-	0.84	0.48	0.48	0.48	0.71	0.79	0.67	0.41	0.84	0.37	0.45
背骨=○:肺呼吸=○:毒=×:ひれ=×:尾=○:家庭向き=×	6	-	-	-	-	-	0.40	0.40	0.53	0.59	0.66	0.77	0.34	0.68	0.43	0.37
卵生=×:哺乳性=○:歯=○:背骨=○:肺呼吸=○:毒=×:タイプ=哺乳類	7	-	-	-	-	-	-	1	0.33	0.51	0.44	0.38	0.67	0.53	0.48	0.95
羽=×:卵生=×:哺乳性=○:歯=○:背骨=○:肺呼吸=○:毒=×:タイプ=哺乳類	8	-	-	-	-	-	-	-	0.33	0.51	0.44	0.38	0.67	0.53	0.48	0.95
家庭向き=×	9	-	-	-	-	-	-	-	-	0.67	0.62	0.69	0.47	0.39	0.45	0.31
背骨=○:毒=×	10	-	-	-	-	-	-	-	-	-	0.90	0.77	0.70	0.59	0.51	0.48
背骨=○:毒=×:尾=○	11	-	-	-	-	-	-	-	-	-	-	0.86	0.59	0.66	0.56	0.42
背骨=○:毒=×:尾=○:家庭向き=×	12	-	-	-	-	-	-	-	-	-	-	-	0.53	0.54	0.66	0.36
羽=×:飛行性=×:歯=○:背骨=○:毒=×	13	-	-	-	-	-	-	-	-	-	-	-	-	0.42	0.73	0.69
水生=×:背骨=○:肺呼吸=○:毒=×:ひれ=×:尾=○	14	-	-	-	-	-	-	-	-	-	-	-	-	-	0.38	0.49
羽=×:飛行性=×:歯=○:背骨=○:毒=×:尾=○:家庭向き=×	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.50
卵生=×:哺乳性=○:飛行性=×:歯=○:背骨=○:肺呼吸=○:毒=×:タイプ=哺乳類	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 9　N-most 法によって選択されたパターン群, および各パターンのペアの冗長率

	パターンのペア (i, j) の冗長率															
パターン	i \ j	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
哺乳性=○:タイプ=哺乳類	1	0	0	0.98	0.98	0.98	0	0	0.95	0.95	0.95	0.46	0.52	0	0.44	0.93
羽=○:タイプ=鳥	2	-	0.34	0	0	0	0.36	0.37	0	0	0	0.26	0	0.37	0.27	0
卵生=○:哺乳性=×	3	-	-	0	0	0	0.90	0.93	0	0	0	0.40	0.32	0.93	0.40	0
卵生=×:タイプ=哺乳類	4	-	-	-	1	1	0	0	0.93	0.93	0.93	0.45	0.51	0	0.42	0.95
卵生=×:哺乳性=○:タイプ=哺乳類	5	-	-	-	-	1	0	0	0.93	0.93	0.93	0.45	0.51	0	0.42	0.95
卵生=×:哺乳性=○	6	-	-	-	-	-	0	0	0.93	0.93	0.93	0.45	0.51	0	0.42	0.95
毛=×:哺乳性=×	7	-	-	-	-	-	-	0.96	0	0	0	0.36	0.36	0.96	0.43	0
毛=×:卵生=○	8	-	-	-	-	-	-	-	0	0	0	0.35	0.34	1	0.42	0
毛=○:タイプ=哺乳類	9	-	-	-	-	-	-	-	-	1	1	0.47	0.49	0	0.41	0.97
毛=○:哺乳性=○:タイプ=哺乳類	10	-	-	-	-	-	-	-	-	-	1	0.47	0.49	0	0.41	0.97
毛=○:哺乳性=○	11	-	-	-	-	-	-	-	-	-	-	0.47	0.49	0	0.41	0.97
肺呼吸=○:ひれ=×	12	-	-	-	-	-	-	-	-	-	-	-	0.48	0.35	0.61	0.46
羽=×:飛行性=×	13	-	-	-	-	-	-	-	-	-	-	-	-	0.34	0.55	0.48
毛=×:卵生=○:哺乳性=×	14	-	-	-	-	-	-	-	-	-	-	-	-	-	0.42	0
背骨=○:尾=○	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.40
毛=○:卵生=×:タイプ=哺乳類	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 10　m-patterns 法によって選択されたパターン群, および各パターンのペアの冗長率