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

Evaluation of an Interestingness Measure of Patterns Based on Frequencies Estimated from Their Subpatterns and Superpatterns

吉田 由起子* 太田 唯子†
小林 健一† 湯上 伸弘†

Yukiko YOSHIDA Yuiko OHTA
Ken'ichi KOBAYASHI Nobuhiro YUGAMI

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. sub+super 法は, 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パターン群を選択することができる. 本稿では, sub+super 法による興味深いパターン群の選択能力を従来手法との比較実験に基づいて評価する.

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. In this paper, we evaluate the ability of the sub+super method to select interesting patterns through a comparative experiment with other existing methods.

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より多く出現する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. 最小サポート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する選択手法を取扱う. 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間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, 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評価対象のパターンの頻度をその部分集合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, それらのアイテム群の部分集合である多数のパターンを同等に興味深いパターン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 sub+super 法を提案している [8, 9]. これは, 評価対象のパターンが包含するサブパターンの頻度, および包含されるスーパーパターンの頻度からパターンの頻度を推定し, パターンの実際の頻度が推定頻度から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手法である. 本手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推測可能なサブパターンに対しては興味深さを低く抑える作用がある. そのため, 本手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 法による興味深いパターン群の選択能力を, 従来手法との比較実験に基づいて評価する.

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. 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を $\mathcal{D}$ で表し, $\mathcal{D}$ 内のトランザクション数を $N_{\mathcal{D}}$ で表す. 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

*正会員 (株) 富士通研究所 y-yoshida@jp.fujitsu.com
†非会員 (株) 富士通研究所 {yuiko, kenichi, yugami}@jp.fujitsu.com

表 1 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
Table 1 Description for Databases and Frequent Pattern Generation in the Experiment

データベース	internet-ads	dna	mushroom	zoo	soybean
トランザクション数	3279	2000	8124	101	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)
アイテムの種類	1557	183	119	43	117
アイテムの頻度の平均	0.00884	0.254	0.193	0.395	0.307
アイテムの頻度の分散	0.00105	0.00301	0.0559	0.0716	0.0904
最小サポート/トランザクション数	0.025	0.05	0.05	0.05	0.3
パターンの最大サイズ	8	5	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

されるパターンを $\boldsymbol{s}^+ \setminus \boldsymbol{s}$ で表す. パターン $\boldsymbol{s}^-$ と $\boldsymbol{s} \setminus \boldsymbol{s}^-$ 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.

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

同様に, $\boldsymbol{s}$ と $\boldsymbol{s}^+ \setminus \boldsymbol{s}$ 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部分集合であるパターン群の多くを興味深いパターン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.

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も用い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. ここで, 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十分に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, 実際の頻度と推定頻度との比の最小値を用いる. また, 評価値を有限値に抑えるために, 実際の頻度と推定頻度との比の arctan を取っている. $1/\pi$ は正規化のための係数である.

3. 実験

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. 実験に用いたデータベースを表 1 に示す. これらは, StatLog [7] および UCI Machine Learning Repository [2] 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のデータベースに変換したものである[1].

3.1 比較対象の手法

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 1 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 $M = 200$ 個を選択して比較した.

sub $I_{sub+super}$ 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

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

の値が大きいアイテム集合を選択する手法. sub+super 法におけるスーパーパターン側の効果を調べるために用いる.

N-most 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 [4]. 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 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 [6]. 相互依存度は, アイテム集合に属する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, $f(\boldsymbol{s}) \geq minsup$ 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.

[1] 機械学習用データベースにおけるブーリアン値や離散値属性のリストを「属性名=値」という形式のアイテムのリストに変換した.

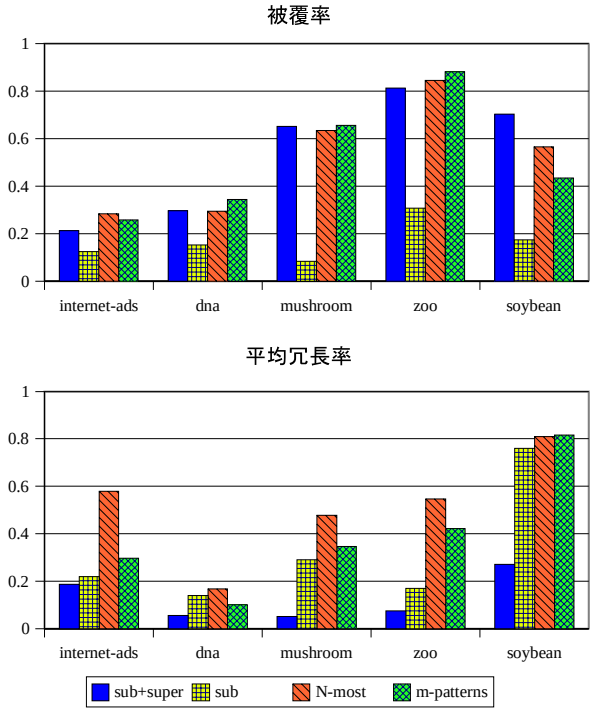

図 1 被覆率および平均冗長率

Fig. 1 Cover and Redundancy Rates

なお, sub+super 法におけるサブパターンおよびスーパーパターンの生成には種々の方法が考えられるが, 今回の実験では, 評価対象のパターンに対してサイズが 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た.

3.2 評価方法

各データベースから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.

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評価するための尺度として, 次式で定義される被覆率を用いる.

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

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. 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.

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評価するための尺度として, 次式で定義される平均冗長率を用いる.

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

表 2 データベース zoo から 3 手法によって選択された興味深いパターン群

Table 2 Interesting Patterns Selected by Three Methods from Database Zoo

	sub+super
1	羽=×, 卵生=○, 哺乳性=×, 毒=○
2	毛=○, 羽=×, 飛行性=○, 水生=×, 捕食性=×, 肺呼吸=○, ひれ=×, ネコの大きさ=×
3	羽=×, 卵生=○, 飛行性=×, 水生=○, ひれ=×, 足=4, 家庭向き=×
4	毛=×, 羽=×, 哺乳性=×, 飛行性=×, 捕食性=○, 毒=○, 家庭向き=×
5	水生=×, 捕食性=×, 背骨=○, 肺呼吸=○, 毒=×, ひれ=×, 尾=○, 家庭向き=○
6	捕食性=×, 背骨=○, 毒=×, 家庭向き=○, ネコの大きさ=×

	N-most
1	毒=×
2	毒=×, 家庭向き=×
3	肺呼吸=○, 毒=×, ひれ=×
4	背骨=○, 肺呼吸=○, 毒=×, ひれ=×
5	背骨=○, 肺呼吸=○, 毒=×, ひれ=×, 尾=○
6	背骨=○, 肺呼吸=○, 毒=×, ひれ=×, 尾=○, 家庭向き=×

	m-patterns
1	哺乳性=○, タイプ=哺乳類
2	羽=○, タイプ=鳥
3	卵生=○, 哺乳性=×
4	卵生=×, タイプ=哺乳類
5	卵生=×, 哺乳性=○, タイプ=哺乳類
6	卵生=×, 哺乳性=○

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の平均である. 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.

3.3 評価

図 1 は各データベースから各手法によって選択されたパターン群による被覆率および平均冗長率である. sub+super を他の手法と比較すると, データベース soybean では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 と同程度かやや低いという結果になった. sub+super 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 で選択されたパターン群は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 で選択されたパターン群は, 他の手法と同程度かそれ以上の被覆率を示した. 平均冗長率については, sub+super は, 他の手法に比べて非常に低い, すなわち優れた値を

示した．また, sub+super 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 は, 被覆率が他の手法に比べてかなり低く, 平均冗長率が sub+super と N-most, m-patterns との間の値を取ることから, sub+super 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．

表 2 は, データベース zoo から sub+super, N-most, m-patterns によってそれぞれ選択された上位 6 パターンを表す．N-most の 1–5 位のパターン群は, 6 位のパターンの部分集合になっている．これは, パターンのサイズ k_{min} から k_{max} までそれぞれ頻度の上位のパターンを選択すると, 特定の頻出パターンの部分集合ばかりが選択されやすくなることを示している．また, m-patterns の 1, 4, 5, 6 位は哺乳類という大きな集合が共通に持つありふれた特徴を述べているに過ぎない．また, 2 位は鳥類についての, 3 位は非哺乳類についてのありふれた特徴である．これは, 部分集合間の相互依存度の高さによってパターンを評価すると, 共起性が高いアイテム群の部分集合が繰り返し選択されたり, ありふれたパターンが選択されやすいことを示している．一方, sub+super のパターン群は, 互いに非常に異なっていて, しかも, 種々の動物の特徴を表すデータベースである zoo の中で意外性のある特徴を表している．これは, sub+super が, データベース内の個々の特徴の出現頻度と比較して意外に多く出現する特徴の組合せのパターン群の中から, 冗長性が低いものを選択しているからである．なお, 他のデータベースについても, zoo と同様の傾向が見られた．

したがって, sub+super は, 他手法よりも冗長性が非常に低く, 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率的に表現できる興味深いパターン群を選択していることが分かる．

4. 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って, sub+super 法による興味深いパターン群の選択能力を評価した．共起性の高いアイテム群が存在するデータベースにおいて, sub+super 法は, 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できるパターン群を選択することを示した．

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.

[9] 吉田, 太田, 小林, 湯上．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．電子情報通信学会第 15 回データ工学ワークショップ *(DEWS)* / 第 2 回日本データベース学会年次大会 予稿集, 2004.

吉田 由起子　Yukiko YOSHIDA
(株) 富士通研究所．1992 東京工業大学大学院修士課程修了．データマイニングの研究・開発に従事．情報処理学会会員．人工知能学会会員．日本データベース学会会員．

太田 唯子　Yuiko OHTA
(株) 富士通研究所．1993 東京工業大学大学院修士課程修了．組合せ最適化の研究・開発に従事．情報処理学会会員．人工知能学会会員．電子情報通信学会会員．

小林 健一　Ken'ichi KOBAYASHI
(株) 富士通研究所．1994 東京大学大学院修士課程修了．コンピュータアーキテクチャ, データマイニング, ソフトウェアエンジニアリングの研究・開発に従事．情報処理学会会員．

湯上 伸弘　Nobuhiro YUGAMI
(株) 富士通研究所．1989 東京工業大学大学院修士課程修了．組合せ最適化, 知識発見等の研究に従事．情報処理学会会員．人工知能学会会員．