F-054

頻度の閾値を考慮してルールの信頼度を推定する手法の提案
Developing a Confidence Estimator of Association Rules Based on Frequency Thresholds

今西 咲季[1)]　菊地 真人[1)]　大囿 忠親[1)]
Saki Imanishi　Masato Kikuchi　Tadachika Ozono

1　はじめに

組み合わせ爆発にともなう計算的・空間的コストの問題から，統計量推定の際に低頻度事象に対する推定を無視することはよく行われている．その一例として，データベースから関連するアイテム集合の組み合わせを網羅的に発見する相関ルールマイニングがある．相関ルールマイニングでは，アイテム集合 A が出現するならば別のアイテム集合 B も出現するという関係を相関ルール A→B で表現し，A→B の関連の強さを条件付き確率 P(B | A) で定義する．P(B | A) は信頼度と呼ばれ，ルールの価値を表す重要な指標となる．信頼度の高いルールほど A，B 間の関連が強いと予測されるため，優先してマイニングされる．なお，ルール生成の際に考えうるアイテム集合の組み合わせを全列挙すると組み合わせ爆発を起こす．そのため，実用上はルールの出現頻度に閾値を設け，閾値よりも頻出するルールに対してのみ信頼度を推定する．これは素朴な対処法であるが，閾値周辺の頻度を持つルールに対する推定値が大きく異なる問題がある．特に，閾値よりもわずかに高い頻度を持つルールの信頼度が過大推定され，偶然に共起しただけの A，B の組み合わせが優先して発見されてしまう．

そこで本研究では，閾値を考慮した信頼度の推定量を相関ルールマイニングに導入する．具体的には，頻度が閾値に近づくにつれて信頼度をあえて保守的に（低めに）見積もるという発想を取り入れる．この推定量は，従来の枠組みをわずかに変更するのみで導入でき，マイニング時の各種コストを増加させない利点がある．本手法の有効性を検証するには，発見されたルールが真の関係を表すか否かが重要になる．しかし一般に，ルールが表す関連性の真偽を明確化することは困難である．そこで相関ルールを用いた分類タスクで評価を行った．本評価では，発見されたルール集合の “品質” を分類精度として定量化できる．結果として，閾値を考慮した推定量を用いると分類効率を維持して，分類精度の減少を高々 7%に抑え，使用するルール数を大幅に削減できた．この結果は，我々の方法が真に関連するルールをより優先的に発見できたことを示唆する．なお本研究では，頻度に閾値を用いる一例として相関ルールマイニングに着目したが，提案手法は他の応用例にも容易に適用できる．

2　一般的な推定量

一般に信頼度は次式で示すように，アイテム集合に対する出現頻度の比を取ることで推定される．

$$\widehat{\mathrm{P}}(\mathrm{B} \mid \mathrm{A}) = \frac{f(\mathrm{A},\mathrm{B})}{f(\mathrm{A})} \quad \text{if } f(\mathrm{A},\mathrm{B}) > \mathrm{minsup} \qquad (1)$$

ここで，$f(\mathrm{A})$ はあるデータベースにおいてアイテム集合 A を含むトランザクションの出現頻度，$f(\mathrm{A},\mathrm{B})$ は A と B をともに含むトランザクションの出現頻度を表す．

1)　名古屋工業大学大学院 知能情報プログラム

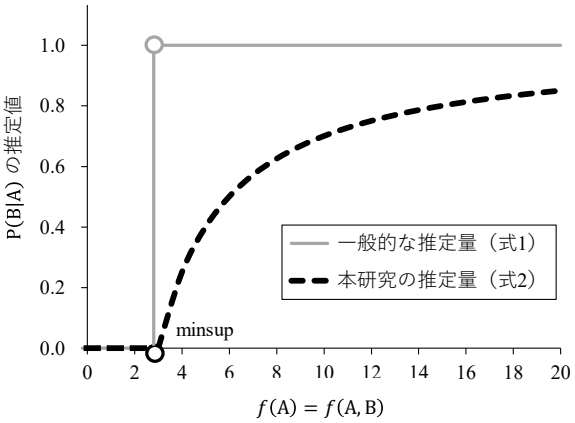

図 1　頻度に対する推定値の変化

minsup は頻度の閾値であり，最小支持度と呼ばれる．上式は信頼度の推定に最小支持度を考慮しないため，最小支持度の前後で推定値が大きく異なる問題がある．この問題の具体例を図 1 に示す．この図は横軸にアイテム集合の頻度を，縦軸にその頻度における信頼度の推定値を取るグラフである．minsup は頻度の閾値であるため，横軸のある一点を取る．図中では理解のために，$f(\mathrm{A})$ と $f(\mathrm{A},\mathrm{B})$ が等しい状況，つまり頻度が最小支持度より大きい場合は常に $\widehat{\mathrm{P}}(\mathrm{B} \mid \mathrm{A}) = 1$ となる状況を仮定している．図 1 に示すように，頻度が閾値より大きいと推定値が常に最大値 1 を取るのに対し，閾値以下になると途端に 0（実際は推定値なし）となってしまう．しかし実際には，わずかな頻度の差で信頼度が大きく異なるとは考えにくく，このような推定値の遷移は不自然である．さらに閾値周辺は低頻度であり，低頻度のルールに最大の信頼度を与えることは，誤った関係の発見につながる．

3　本研究で使用する推定量

本研究では，我々が先行研究 [1, 2] において導出した次式の推定量を使用する．

$$\widetilde{\mathrm{P}}(\mathrm{B} \mid \mathrm{A}) = \frac{f(\mathrm{A},\mathrm{B}) - \mathrm{minsup}}{f(\mathrm{A})} \quad \text{if } f(\mathrm{A},\mathrm{B}) > \mathrm{minsup} \qquad (2)$$

この推定量は，信頼度の真値と推定モデルの二乗誤差を最小化する理論的枠組みで導出された．そして図 1 から分かるように，最小支持度に近づくにつれて推定値が保守的に見積もられるという挙動を実現している．また得られた式は，式 (1) の分子から最小支持度を減算するという単純なものである．したがって，相関ルールマイニングの既存アルゴリズムに容易に導入できる実用性を持つ．さらに追加の計算をほぼ必要としないため，マイニング時に重要となる計算時間やメモリ使用量といった各種コストの増加を抑えることができる．

表 1　分類性能の比較結果

データ	ルール数↓		精度 [%]↑		計算時間 [s]↓	
セット	従来	提案	従来	提案	従来	提案
balance	39.00	15.70	84.50	77.14	0.15	0.13
breast-w	17.60	5.80	96.04	94.71	13.54	11.54
ecoli	11.40	9.60	80.02	76.69	0.64	0.56
glass	12.30	11.60	70.29	67.43	18.45	14.94
iris	4.90	3.60	94.67	94.67	0.02	0.02
lenses	1.60	1.70	66.67	71.67	0.01	0.01
mammo	33.00	11.90	81.17	83.05	0.41	0.32

4　評価実験

式 (2) の有効性を検証するため，相関ルールマイニングで得たルール集合の品質，すなわちルール集合にある真の関係を持つルールの割合を評価したい．しかし一般にマイニング手法の評価軸は，計算時間や発見したルール数であり，ルールが表す関係が真か否かを定量的に評価することが難しい．そこでルール集合を用いる分類問題を解き，分類に用いたルール数，分類精度，計算時間で有効性を評価する．このとき，分類に使用されるルールは $condset \rightarrow l$ と定義される．ここで $condset$ は分類の根拠となる属性値の集合，l は予想されるクラスラベルである．本タスクでの “良いルール” とは，多くの分類対象を正確に分類できるルールである．そして，良いルールのみからなる高品質なルール集合を得るマイニング手法を優れているとみなす．高品質なルール集合では，少ないルールでデータセットの分類規則を正確に示すことができる．よって使用されるルール数は少なく，分類精度は高く，計算時間は短くなる．本研究では相関ルールによる最も基本的な分類器 CBA[3] を使用する．これは単純な分類器を用いると，信頼度の推定以外が分類性能に及ぼす影響を極力排除できると考えたためである．CBA の変更箇所は信頼度の推定に式 (1) を用いるか式 (2) を用いるかのみである．その他の CBA の設定は，CBA の元論文と同様である．UCI リポジトリ[1)] にある 7 種類のデータセットを分類に用いた．

各データセットの分類に使用したルール数，分類精度，計算時間を表 1 に示す．ここで計算時間は分類器の訓練に要した時間である．分類器の使用するルールを選ぶ基準として式 (1) を用いた場合を従来，式 (2) を用いた場合と提案と表記している．ルール数と計算時間は少ないほど，分類精度は高いほど良く，各値は 10 分割交差検証の結果である．各データセットにおいて，二手法のうち良好な性能に下線を引いて強調した．結果として，式 (2) を用いるとルール数はいくつかのデータセットで大きく減少した．これは，低頻出なルールの信頼度が式 (1) よりも低めに見積もられ，高頻出かつ信頼度も高い有用なルールが分類に多く用いられたためと考える．また使用されるルールが少なくなったことにともない，訓練に要する時間も短くなったことが分かる．その一方で lenses と mammo を除くデータセットでは，分類精度は式 (1) を用いた場合と同等または劣るが，精度の減少を最大でも 7%ほどに抑えることができた．

1) https://archive.ics.uci.edu/

5　考察

実験結果から，最小支持度を考慮した推定量により，ある程度の分類精度を担保しながら使用するルール数を大幅に削減し，計算時間も短縮できたことが分かる．CBA の約 2/3 のルール数で同程度の分類精度を実現しており，解釈性に優れているとも言える．一方，分類精度はただの CBA に劣る傾向があり，改善の必要性も示唆された．ここでは精度低下の原因について考察する．

精度低下の主要因を二点示す．一点目は，使用したデータセットによる分類が簡単すぎる点である．分類結果を精査したところ，用いた多くのデータセットでは，低頻出なルールでも分類対象を精度よく分類できた．一方，提案手法が精度で勝る mammo では，低頻出なルールにそれなりの誤りが混入していた．我々は当初，低頻出なルールは偶然に出現したアイテム集合から構成されるため，分類に使用すると誤った分類結果を導くと考えていた．そのようなルールは誤った関係を誘発するため，実際に相関ルールマイニングでは問題視されてきた．しかし，使用したデータセットによる傾向は，我々の前提とは一致しなかった．そのため，低頻出のルールは不正確で誤りが混ざるという考えのもとで，信頼度を低めに見積もる作用が精度の低下を招いたと考察する．

考えられる原因の二点目は，信頼度を低めに見積もる度合いを，提案の推定量が柔軟に制御できない点である．データに混在する誤りの量や分類器で設定した各種の閾値に応じて，信頼度を低めに見積もる度合いは最適化されるべきと考える．しかし式 (2) から分かるように，現在の手法にはそのような機能はない．よってマイニングの効率を落とさず，信頼度の推定をより柔軟に制御する方法の考案が今後の課題である．

6　おわりに

本研究では，最小支持度を考慮した信頼度の推定量を提案した．本手法は最小支持度を推定式に取り入れ，ルールの頻度に応じて信頼度を保守的に見積もる．提案手法を相関ルールに基づく分類器 CBA に導入して有効性を検証した．そして，分類精度の減少を高々 7%程度に抑えつつ，ルール数と計算時間を削減できることを示した．今後の課題として，精度についての利点や改善点を探るため，詳細な検証や手法の再考が必要である．また，高度な分類器へ提案の推定量を導入し，効果を検証したい．なお本研究では，頻度に閾値を利用する一例として相関ルールマイニングに着目したが，提案手法は他の応用例にも容易に適用できる．

謝辞

本研究の一部は JSPS 科研費 JP22K18006，JP24K03052 の助成を受けたものです．

参考文献

[1] T. Aoba, M. Kikuchi, M. Yoshida, K. Umemura, “Improving association rule mining for infrequent items using direct importance estimation,” In Proc. ICAICTA’20, pp.1-5, 2020.

[2] 日下部友飛, 菊地真人, 大囿忠親, “Associative Classifier におけるソフトな最小支持度の導入,” DEIM2023, pp.1-8, 2023.

[3] B. Liu, W. Hsu, Y. Ma, “Integrating classification and association rule mining,” In Proc. KDD’98, pp.80-86, 1998.