F-044

系列パターンマイニングのためのカスケードモデル
Cascade model extension to sequential pattern mining

吉川　芳浩　　岡田　孝
Yoshihiro Yoshikawa　　Takashi Okada

1. はじめに

近年，様々な分野で大量のデータから有用な知識を発掘するデータマイニングが行われている．本研究室ではクラス相関ルールの拡張であるカスケードモデルを提案している[1][2]．これは特徴的ルール導出の一種であり，理解を目的としたマイニングでの有効性が示されている．この方法はアイテムセットのラティスを構築し，群間平方和(BSS: Between groups sum of squares)を用いてリンクの識別力を定量化する．それにより，ラティス内で識別力が大きいリンクをルールとして提示する．また識別力は単一パラメータで提示され，これによりルールを最適化するため，出力されるルール数は相関ルールに比べて 1/10～1/100 と少なくなる．

本研究では系列データを説明要因とし，クラス属性を識別できるようにカスケードモデルを拡張する，頻出の系列パターン導出には系列パターンマイニングの代表的なアルゴリズムである PrefixSpan[3]を用い，得られたパターン間で急激なクラス属性の分布変化を検出して，ルールとして表現するような，新しい系列パターンマイニング法を提案する．このシステムを利用しオンラインショップの Web 閲覧履歴から，購入者と非購入者を識別できるような特徴的パターンの発見を試みる．

2. 特徴的な系列パターン

2.1 ラティス作成

表 1 に示すサンプルデータは，ID とクラス属性 Z および Sequence から成り立っている．Sequence 内の< a >, < b >をアイテムと呼び，アイテムの集合 < a >, < (a, b) >, < b >をエレメントと呼ぶ．同時に共起する複数のアイテムからなるエレメントは，"( )"で囲んで表記する．Sequence はエレメントの系列である．

表 1 サンプルデータ

Sid	Z	Sequence
1	y	a,(a,b),b
2	y	a,b
3	n	b,a
4	y	b,b
5	n	a,b
6	y	(a,b)
7	y	a,(a,b)
8	n	a,b

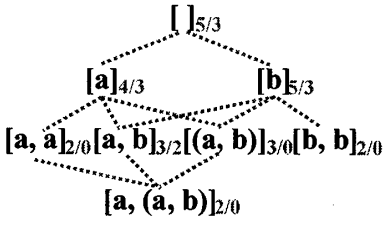

図 1 系列パターンのラティス表現

表 1 のサンプルデータから，頻出の系列パターンを導出し，パターン間のサブシーケンス関係に従い図 1 のようにラティスを生成する．ここでは最小支持度を 2 としている．

関西学院大学大学院 理工学研究科 情報科学専攻
Department of Informatics, Graduate School of Science and Technology, Kwansei Gakuin University

図 1 のラティスを見ると，パターン[a]からパターン[(a, b)]へ移る際に，Z の分布が(y/n): 4/3 ⇒ 3/0 と大きく変化しており，特徴的なルールとして表現する価値のあることが判る．

2.2 ルール表現

従来のカスケードモデルでルールは次のように表されている．

```
IF [a2] ADDED ON []
   THEN Z(y/n):    5/3 ⇒ 5/0 BSS=0.70
   Then B(b1/b2):  5/3 ⇒ 4/1 BSS=0.15
```

このルールには，前提条件[ ]に主条件[a2]が付加される時，目的変数 Z の分布変化とそれに伴う BSS 値が記されている．さらに主条件[a2]に伴い同時に大きな分布変化を示す説明変数 B の情報も付加相関として記されている．

本研究では，系列パターンを対象としたルールに次の形式を提案する．

```
IF 拡張条件 DERIVED FROM 前提条件
   THEN 目的変数: 分布変化　BSS 値
   Then 系列パターン: 分布変化　BSS 値
```

図 2 の[a] ⇒ [(a, b)]にこの表現を適用すると下記のようになる．

```
Rule 1:
  IF [(a, b)] DERIVED FROM [a]
    THEN Z(y/n): 4/3 ⇒ 3/0  BSS=0.55
    Then [a, a]:   2/5 ⇒ 2/1  BSS=0.44
```

カスケードモデルの特徴の一つとして付加相関が挙げられる．付加相関とは，主条件と相関の高い説明変数を，付加的な右辺情報としてルール中に表示するものであった．付加相関により，ユーザーはルールの周辺状況を見極め，クラス属性が識別される本当の原因を知ることが出来る．系列パターンマイニングにおいても，パターン[a, a]の出現割合が，前提条件[a]の元では 7 事例中 2 事例のみ現れるのに対し，拡張条件[(a, b)]の元では 3 事例中の 2 事例中と大きく変化しており，付加相関として提示する価値があろう．実際，ラティス最下部のパターン[a, (a, b)]が本質的な役割を果たしているとすれば，このルールからそれを推定できることになる．この結果は，拡張条件だけが識別力の高い系列パターンではない可能性を示唆している．

ルール識別力の表現は，式(1)に示す群間平方和（BSS: Between-groups sum of squares）を用いる．ただし，n は支持事例数を表し，p(a)はその属性が値 a を取る確率である．また添え字 U, L は，ラティスの節点の上側(upper node)と下側(lower node)に対応する．

$$BSS_i^g = \frac{n^L}{2}\sum_a \left(p_i^L(a) - p_i^U(a)\right)^2 \qquad (1)$$

3. アルゴリズム

3.1 PrefixSpan

Pei らは，系列パターンマイニングに対する効率的な手法として PrefixSpan を提案している．PrefixSpan は候補シーケンスを作成せず，射影(Prefix-projection)と呼ばれる操作を，深さ優先で再帰的に行うことで，頻出の系列パターンを抽出するアルゴリズムである．射影とは，元の系列データから射影対象の系列(Prefix)より後ろに存在する系列(Postfix)のみを抽出する操作である．

ここで，表 1 のサンプルデータに PrefixSpan アルゴリズムを最小支持度 2 で適用すると，図 2 のラティスに示す 7 つの頻出系列パターンが出力される．従来の PrefixSpan では木構造は作成されないが，本研究では頻出パターンが導出されるごとにリンクを張る．この操作で繋がれたリンクを図 2 上では，実線で表現した．さらに，ノードにはクラス属性の分布と支持レコードの ID を記憶させる．図 2 ではクラス属性の分布をノード右に表示した．

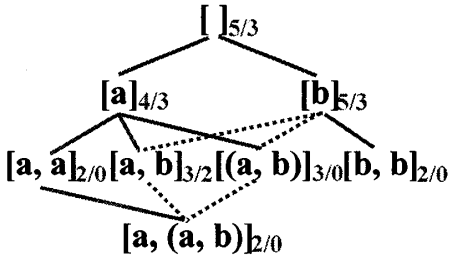

図 2　サンプルデータから作成されるラティス

3.2 幅優先探索によるリンク発見

ラティスを完成させるために，図 2 の点線で表現されているリンクを発見する必要がある．そこで，幅優先で長さ i の系列パターン P_i を調べ，長さ $i-1$ の系列パターン Q_{i-1} で $P_i \subset Q_{i-1}$ となる Q_{i-1} が存在すれば，これらの間にリンクを張る．

3.3 ルール導出

ラティス内のすべてのリンクに対し BSS 値を計算し，図 3 に示すリンクに付した．ユーザーが指定する閾値 (threshold×事例数)よりも高い BSS 値のリンクをルールとして採用する．系列パターン Q_{i-1} と P_i 間のリンクがルールとして採用された時，Q_{i-1} とその子となるノード R_i に付された支持レコード ID を用いて，付加相関を計算する．図 3 に破線で示したリンクをルールとして表現する時，付加相関の候補となる系列パターンは[a, a] , [a, b]の 2 種である．

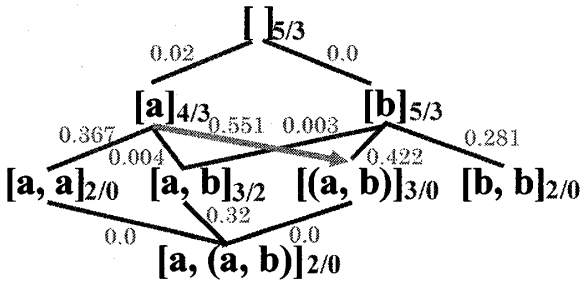

図 3　リンクに BSS 値を付したラティス

4. 実験

対象データとして，著者の 1 人が運用するオンラインショップの情報を参考に Web 閲覧履歴を人工的に作成した．オンラインショップの購買率は 1～2%が実状と言われている中で，店舗の課題は，購買意識の高い訪問者をどれだけ顧客にするかである．実店舗よりはるかにウィンドウショッピングを目的とする訪問者が多く，その差別化を図る事の重要性は高い．

データはエレメント長が 2 以上の系列データを 100 レコード作成した．アイテムの内容は，ホームページ(top)，カテゴリーページ(4 種類)，商品ページ(10 種類)，名入れページ(naire)である．なお，名入れページは今回扱う全ての商品ページからリンクが張られており，商品の購入を検討する際には閲覧される可能性の極めて高いページであるため，商品ページとまとめて 1 エレメントとする．またクラス属性(purchase) には購入の有無を与える．

最小支持度を 5%，threshold を 1%としてルール導出を行った．得られたルールの中で興味深いと思われるものを次に示す．

Rule 2:
IF [{ naire , itemA }] DERIVED FROM [itemA]
THEN [purchase] (y/n) :11/12 ⇒ 8/1 BSS= 1.5175
Then [itemA, itemB] : 7/16 ⇒ 5/4 BSS= 0.5679

Rule 2 から，商品番号 itemA だけを閲覧した場合より，同時に名入れページを閲覧した方が購買に至る可能性が飛躍的に高まることが読み取れる．更に付加相関より，itemA の後に itemB ページへ移動した場合も少し購買意欲の高い閲覧者であると考えられる．

5. まとめと今後の課題

本研究では，クラスの識別力の高いルールを出力するカスケードモデルの特性を生かして，系列データからの系列パターンルール導出システムを開発した．またサンプルデータではあるが，購買意識の高い閲覧行動をいくつか見つけることが出来た．今後は，実データを利用した解析を行い，有用な利用法を発見していく予定である．

なお，本稿に述べたシステムで，Rule 1 で[b, b]との付加相関は，計算量の増加を避けるために評価していない．しかし原理的にはこのような系列パターンが重要となる可能性がある．今後，高速のアルゴリズムを実装することにより，ルール選択や全ての付加相関計算の機能を取り入れる予定である．

参考文献

[1]Takashi Okada, “Rule Induction in Cascade Model based on Sum of Square Decomposition”, Principles of Data Mining and Knowledge Discovery, PKDD 468-475 (1999).
[2]Takashi Okada, “Efficient Detection of Local Interactions in Cascade Model”, PAKDD, 193-203 (2000)
[3]J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, M. -C. Hsu., “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Prejected Pattern Growth.”, ICDE, 215-226 (2001).