アイテム探索を利用した相関ルールマイニング

田中 靖士† 高野 義士† 杉村 博‡ 松本 一教†‡
†神奈川工科大学情報学部情報工学科 ‡神奈川工科大学大学院情報工学専攻

1 はじめに

データマイニングは，大量にあるデータを網羅的に解析することで知識を取り出す技術であり，時系列データの解析はデータマイニングの重要なテーマである [2]. 時系列データを対象として相関ルールをマイニングする研究にはさまざまな問題がある [2, 3]. 通常のマイニングでは，アイテムやトランザクションがデータを構成する要素となっているが，時系列データの場合，それらがあらかじめ定まっているわけけではない. そのため，時系列データからアイテムやトランザクションを抽出する方法も重要な問題である. 本論文では，時系列データを対象として相関ルールを発見するデータマイニングシステムについて報告する.

2 相関ルールマイニング

トランザクションからなるデータベースを考える. 各トランザクションはアイテムから構成される. 商店における販売履歴を例にすれば，1回の買い物が1トランザクションであり，1回の買い物の中に含まれる複数の品物がアイテムである.

Aをアイテム集合とするとき，Aを含むトランザクション集合を$T[A]$と書く. 集合の要素数を示すには$\sharp$を使用して$\sharp T[A]$と書く. 共通要素を含まないアイテム集合XとY $(X \cap Y = \phi)$に対し，相関ルール$(X \rightarrow Y)$の支持度s(support)，信頼度c(confidence)は次で定義される. ただし，Nはデータベース中の全トランザクション数である.

$$s = \frac{\sharp T[X \cup Y]}{N}, c = \frac{\sharp T[X \cup Y]}{\sharp T[X]}$$

与えられた最少支持度$s_{\min}$に対し，それ以上の支持度を持つアイテム集合Xを頻出アイテム集合という. Xの要素数がkのとき，サイズkの頻出アイテム集合という. 相関ルールマイニングは，次の2ステップに分けて実行する.

1. トランザクションデータベース全体を探索し，すべてのサイズの頻出アイテム集合全体Fを求める.
2. 上記ステップで求めたFを探索し，与えられた最少信頼度以上のすべての相関ルールを生成する.

3 時系列データマイニング

時系列データを対象とした相関ルールマイニングを行うには，時系列データをアイテムからなるトランザクションとして扱う必要がある. 本章ではその方法を説明する.

3.1 時系列データのトランザクション化

時系列データを区切ってトランザクション化する方法は以下の通りである.

1. 時系列データを$s = (x_1, x_2, \cdots, x_n)$とする.
2. 元の時系列データsを部分時系列データS_iに分割する. ただしここでは，各S_iのサイズを等しくしている. すなわち$s = (S_1, S_2, \cdots)$となる. 各S_iがアイテムの候補となるが，そのままではわずかな違いも区別されているので，アイテムとしては不適当である. そこで次節で説明するDTWによるアイテム化を実行し，その結果を$l(S_i)$とする. これらがアイテムとなる.
3. あるkに対して，連続するk個のアイテム毎に区切ったものがトランザクションとなる. $((l(S_i), l(S_2), \cdots, l(S_k)), (l(S_{k+1}), \cdots, l(S_2k)), \cdots)$
4. 与えられた時系列データの終端x_nを含むすべての部分時系列データをトランザクション化する.

時系列データに対するトランザクションと部分時系列データの関係を図1に示す.

3.2 DTW(Dynamic Time Warping)

DTWとは，パターンの要素間に定義された類似度にもとづいて，パターンの伸縮まで考慮に入れたマッチング方式である [1]. この方式ではいくつかのパラメータを事前に定めて，パターンの類似度$s(p1, p2)$を計算できる. ここで，$p1$と$p2$の長さが異なっていても構

Association Rule Mining using Item Search

†Yasushi TANAKA,†Yoshiaki TAKANO,
‡Hiroshi SUGIMURA,††Kazunori MATSUMOTO
†Department of Information and Computer Sciences, Kanagawa Institute of Technology,‡Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology
1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN
{s055101,s055090}@cce.kanagawa-it.ac.jp,
hiroshi.sugimura@gmail.com, matumoto@ic.kanagawa-it.ac.jp

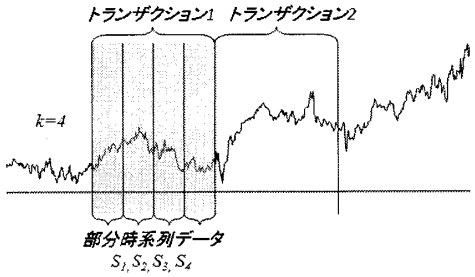

図 1: トランザクション化と部分時系列データ

わない．与えられた実数値 δ に対して $s(p1, p2) < \delta$ であれば，$p1$ と $p2$ が一致すると判定する．δ のことを許容相異度と呼ぶ．このようにして，時系列上のパターンを比較する方法を定めて実装できる．

3.3 部分時系列データのアイテム化

部分時系列データをそのままトランザクションのアイテムとすると，アイテムの種類が膨大となり，良い相関関係を得ることができない．似通った部分時系列データをまとめ，分類したものをトランザクションのアイテムとする．本論文では，この分類化に DTW を用いることを提案する．これにより，部分時系列データが膨大な量になったとしても許容相異度を与えることで動的にアイテム化でき，比較的高速なシステムを維持できる．さらに DTW に与える部分時系列データは，前日からの変化率を用いる．これは株価のように基本値が定まらないような時系列データに有効である．株価は銘柄ごとに取引単位株数が違うため，銘柄が異なると株価の基本値も異なる．このように基本値の異なる時系列データは，元の時系列データからデータの変化率を用いた時系列データを作成し，その変化率による時系列データで処理を行うことで汎用的なシステムとなる．

4 実験

典型的な時系列データとして，ある会社の 10 年間の株価の終値を用いる．1 週間 (5 日) 間隔で区切ったものを 1 アイテムとし，1ヵ月 (4 アイテム) を 1 トランザクションとした．その他のシステム設定は表 1 に示す．実験の方法は許容相異度を 5 から 20 の間で変化させ，発見される相関ルール数の評価を行った．

評価結果を表 2 に示す．アイテム種数とは，DTW でアイテムを分類した分類数である．

表 1: 評価設定

設定名	設定値
アイテム時系列データ数	5
トランザクションアイテム数	4
最少支持度	0.01
最少信頼度	0.50

表 2: 評価結果

許容相異度	アイテム種数	1:1 相関	2:1 相関
5	629	0	0
10	570	0	0
15	368	4	0
20	143	7	17

5 おわりに

本論文では，時系列データ内の相関ルールをマイニングするシステムを提案した．相関ルールを発掘するための最適なパラメータを見つけ出すことは困難である．中でも許容相異度の決定は困難であり，値を大きくすればアイテム数が少なくなり，得られる相関ルールの数が増加する．しかし，あまりに許容相異度を大きくすると，似ていないアイテムを同じアイテムとして扱うことになるため，得られた相関ルール自体価値が減少する．逆に，許容相異度の値を小さくすれば，アイテム数が多くなり相関ルールを得ることができない．

また，トランザクションの大きさを変更して導き出された相関ルールは，異なる意味をもつと思われる．本論文では，1 週間を 1 アイテム，1 か月を 1 トランザクションとしたが，より大きな区切りによってマイニングされるルールは広域的な視点によるルールといえる．

参考文献

[1] Donald J. Berndt, James Clifford, Finding Patterns in Time Series : A Dynamic Programming Approach, Advances in Knowledge Discovery and Data Mining, pp.229-248, 1996.

[2] M. Last, A. Kandel, H. Bunke(edt), Data Mining in Time Series Databases, World Scientific, 2004.

[3] 杉村 博, 松本 一教, DP マッチングを利用する時系列データからのデータマイニング, 第 22 回 人工知能学会 全国大会論文集, 2008.