2H2-04

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

Datamining of Time Series Data based on DP matching

杉村 博　　松本 一教
Hiroshi SUGIMURA　Kazunori MATSUMOTO

神奈川工科大学大学院 情報工学専攻
Graduate Shool of Kanagawa Institute of Technology

This paper presents a data mining method of association rules from time series databases. The usual association rules are extracted by investigating co-occurence patterns of items in a transaction. This direct approach, however, cannot applicable in the case of time series database. There could be many different co-occurence patterns of items in a typical time series database. Thus we often fail to find useful patterns because of minor differences. In order to solve this problem, we identify, by discarding unessential features, a set of similar items as an abstract item. This identification process is carried out based on DP matching with similarity measures.

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 [1]，[3]．時系列データとは，毎日の気温や視聴率，商品販売数など，時間軸で連続しているデータのことで，これらデータを解析することは，未来予測や市場調査を行う上で重要である．一方，データ中のパターンの同時生起に注目した相関ルールをマイニングする方法の有効性が知られている．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もあるが [1]，さまざまな問題がある．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る．時系列データを対象とする場合には，それらがあらかじめ定まっているのではないため，元のデータからどのようにしてアイテムやトランザクションを抽出するか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. DP マッチング

DP(Dynamic Programming) マッチングとは，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 [2].

3. 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．たとえば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 回の買い物のなかに含まれる複数の品物がアイテムである．

相関ルールマイニングでは，このデータベースを解析することによりアイテム間の相関関係をマイニングすることを目的とする．先の商店の例をとれば,「酒を購入した人はお菓子も購入している」というような関係をマイニングすることができる．

連絡先: 杉村博，神奈川工科大学 情報工学専攻 松本研究室，神奈川県厚木市下荻野 1030，sugimura@mm.cs.kanagawa-it.ac.jp

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．

4. 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．

4.1 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. 与えられた時系列データを $s = (x_1, x_2, \cdots, x_n)$ 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 DP マッチング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 を含むまでの，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．

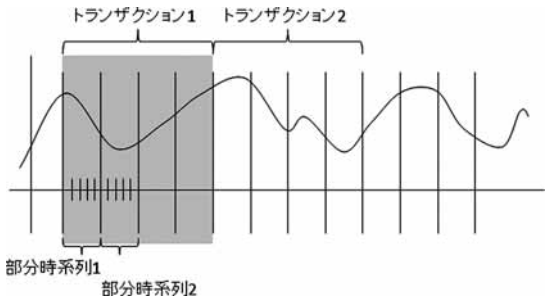

図 1: トランザクション化と部分時系列データ

このような方法によって，時系列データからアイテムを見出し，それらを要素とするトランザクションに分割することができる．

4.2 部分時系列データのアイテム化

時系列データを単に区切った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，DP マッチングを用いることを提案する．

DP マッチングを用いて，部分時系列データの類似度での分類を行う．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類似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 DP マッチング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もち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．DP マッチングによるアイテム化を図 2 に示す．

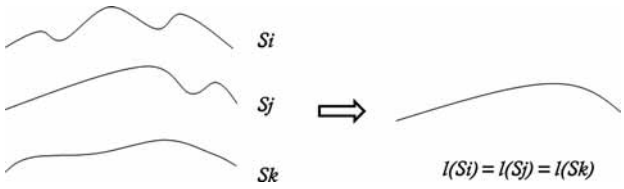

図 2: アイテム化

5. 実験

典型的な時系列データとして，ある 1 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 か月 (4 アイテム) を 1 トランザクションとした．

時系列データは CSV ファイルとして与える．その他のシステム設定は表 1 に示す．ここで，DP マッチング類似度とは，2 つのアイテムをマッチングにより同一視するレーベンシュタイン距離の閾値である．

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
DP マッチング類似度	20
最少支持度	0.01
最少信頼度	0.50

評価結果を表 2 に示す．1:1 相関とは $Item94 \Rightarrow Item2$ のような，アイテム 1 つに対してアイテムが 1 つ相関関係をもつルールのことである．2:1 相関とは $Item25 \cap Item28 \Rightarrow Item4$ といったような，アイテム 2 つに対してアイテムが 1 つ相関関係をもつルールのことである．

表 2: 評価結果

結果名	値
1:1 相関発掘数	22
2:1 相関発掘数	19

6. 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．

相関ルールをマイニング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．なかでも DP マッチングの類似度決定は困難であり，許容する類似度を大きくすればアイテム数が少なくなり，相関ルールの数が増加する．しかし，あまりに許容する類似度を大きくすると，得られた相関ルール自体の価値が減少する．

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．たとえば本論文では 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] M. Last, A. Kandel, H. Bunke(edt): Data Mining In Time Series Databases, World Scientific, 2004.

[2] 大桃諭，陳漢雄，古瀬一隆，大保信夫: タイムワーピングに基づく時系列データの類似検索 - 次元縮小による効率化，DBSJ Letters Vol.4, No.1, pp.1-4.

[3] 安田征吾，廣川佐千男: 短期・中期移動平均線を用いた株価の解析，情報処理学会研究報告，2005-MPS-54(6), Vol.2005, No.37, pp.23-26, 2005.