3D-2

テンプレートにもとづく時系列データからの相関ルールマイニングシステム

杉村博† 高野義士‡ 田中靖士‡ 松本一教†‡
†神奈川工科大学大学院情報工学専攻 ‡神奈川工科大学情報学部情報工学科

1 はじめに

相関ルールマイニングを用いた知識発見の技術は，すでに多くの場面で適用されているが，解決すべき課題も残されている [1, 2]．大量に発見されるルールから，真に有効な知識のみを取り出す技術開発もその１つである．本研究の特徴は２つある．第１には，発見すべき知識の一部をユーザがヒントとして与え，それに関係する相関ルールのみを選択的に発見するようにして，発見されるルール数の爆発を防ぐ方法を提供している．次に，本研究は，株価などの時系列データを対象としているため，時系列データ上でパターンを取り扱う技術を開発し，それに基づく相関ルールマイニングを実現している．

2 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．

2.1 手書き入力グラフのパターン化

マイニングはグラフを手書き入力したパターンを用いて行う．ユーザはポインティングデバイスを用いてグラフを記述する．このグラフをサンプリングするように等間隔に数値化する．グラフの分割数が大きいほど詳細に数値化し，分割数が小さいほど荒く数値化する．ここで得られた数値の１つ１つをアイテムとする．

An association rule mining system of time-series data with templates
†Hiroshi SUGIMURA, ‡Yoshiaki TAKANO, ‡Yasushi TANAKA, and †‡Kazunori MATSUMOTO
†Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology
‡Department of Information and Computer Sciences, Kanagawa Institute of Technology
1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN
hiroshi.sugimura@gmail.com, s055101@cce.kanagawa-it.ac.jp, s055090@cce.kanagawa-it.ac.jp, matumoto@ic.kanagawa-it.ac.jp

2.2 時系列データの一般化

手書きグラフとのマッチング前に，時系列データの一般化を行う．

時系列データの種類はさまざまで，それぞれ基準となる数値が異なる．また同種類の時系列データを取ってみても基準となる数値が異なる場合がある．たとえば株価を用いた場合，基準は単位株ごとの値段となっている．1000 株で１単位とする株と，１株で１単位とする株では値段の差があまりにも開いている．このようなデータに対して同じパターンデータを用いるためには，共通の指標をもとにデータを一般化する必要がある．

本論文では，前日からの変化率を一般化データとする．元の時系列データを，変化率による時系列データに変換することで汎用的なシステムとなる．

2.3 時系列データの分割

手書き入力グラフによってマイニングを行う際に，時系列データから部分時系列データを抽出する．まず，抽出サイズとスキップ量というパラメータを与える．抽出サイズの値だけの，連続した時系列データを抽出し，部分時系列データとする．そして部分時系列データと手書き入力グラフでマッチングを行う．その後，次の部分時系列データはスキップ量だけずらして抽出する．部分時系列データは数式１で求められ，マッチング計算回数は数式２で求められる．

$$I_i = \left\lfloor \frac{s_i}{s_{i-1}} \times 100 \right\rfloor \tag{1}$$

$$M = \frac{record - size - 1}{skip} \tag{2}$$

2.4 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 [3]．マッチングの際に必要なプロパティは 3 つあり，

それぞれ，ずれコスト，不一致コスト，許容類似度である．

不一致ペナルティは，不一致コスト d に不一致度をかけたもので，数式3で求められる．不一致度とは，アイテムとアイテムの違いを示すもので，値が大きいほど不一致だとみなす．

$$D = \begin{cases} \log_{10} \left| \frac{I_t}{I_{t-1}} - \frac{J_u}{J_u - 1} \right| \times d & \frac{I_t}{I_t - 1} \neq \frac{J_u}{J_u - 1} \\ 0 & \frac{I_t}{I_t - 1} = \frac{J_u}{J_u - 1} \end{cases} \quad (3)$$

不一致度の計算に対数を用いることで，ずれペナルティ範囲の増大化を防ぐ．また，全不一致の類似度最大値 $L_{\max}$ は数式4で求められる．p は分割数，l は抽出サイズである．

$$L_{\max} = \begin{cases} D_{all} = D \times p \\ S_{all} = s \times p \times l \end{cases} \quad (4)$$

2.5 決定木作成前フィルタ

決定木は，リスクマネジメントなどの決定理論の分野において決定を行う為のグラフであり，計画を立案して目標に到達するために用いられる．決定木は，意志決定を助けることを目的として作られる．決定木の1つに，質問の答えが2つである2分決定木がある．

決定木は，あまり雑多な数値を使用してしまうと，発散し良い結果が得られないため，本研究では取り出した時系列の初期値を100としたときの割合を使用する．このとき，小数点は切り捨てする．計算式を数式5に示す．

$$N_t = \left\lfloor \frac{x_t}{x_1} \times 100 \right\rfloor \quad (5)$$

3 実験

3.1 実験方法

本論文では典型的な時系列データとして，ある数社の10年間の株価の終値を用いる．時系列データはCSVファイルとして与える．

本論文にて作成したシステムを用いて得られたデータを基に，wekaを用いて決定木を作成した．決定木は2分決定木とし，「買いの場合」には yes,「買いではない場合」を no として学習した．

「買いの場合」とは，検索結果の時系列データの初期値よりも高い値が，検索結果最終日から5日以内に1つ以上存在していた場合と定義した．

3.2 評価結果

各銘柄ごとの実際のデータ例を表1に示す．series は抽出した部分時系列データ総数，hit は検索ヒット数，hit%は部分時系列データの総数に対する検索ヒット率，Sott は決定木サイズ，CCI は決定木精度である．

表 1: 検索数の差の例

stock No	series	hit	hit(%)	Sott	CCI(%)
1	329	103	31.3	15	72.8
2	329	136	41.3	5	63.4
6	329	127	38.6	33	58.3
13	329	132	40.1	7	90.9
17	317	118	37.2	17	69.5
average	326	121	37.1	12	75.5

4 おわりに

本論文では時系列データから類似時系列データを発掘するシステムを開発した．

知識を持ったユーザの手書き入力によるテンプレートを与えることで，関係する部分時系列データのみを得ることができ，得られた部分時系列データから相関ルールを発見することができた．ここで発見したルールは，かなりの割合で視覚的に理解しやすい分類木を作成することに成功した．また，作成した分類木は高い精度であることを示した．

今後の研究としては，本システムによって得られた相関ルールを用いて，利益率などの試用評価を行いながら問題点を洗い出し，改良を行うことがあげられる．

参考文献

[1] Oomomo, S., Chen, H., Furuse, K. and Ohbo, N.: Efficient Search of Similar Time Series under Time Warping with Dimensionality Reduction, Vol. 4, No. 1, pp. 17–20 (2005).

[2] Sugimura, H. and Matsumoto, K.: Datamining of Time Series Data based on DP matching, in *The 22nd Annual Conference of the Japanese Society for Artificial Intelligence* (2008).

[3] Last, M., Kandel, A. and Bunke, H.: *Data Mining In Time Series Databases*, World Scientific (2004).