

4AH-5

2段階解析法を用いた 時系列データマイニングシステム*

北島伸克 谷川哲司†

NEC ヒューマンメディア研究所‡

1 はじめに

「データマイニング」とは、収集したデータから「役立つ情報」を発見する技術である。時系列データマイニングにおける「役立つ情報」の1つに、「近い将来に特定の変動（上昇、下降等）が予測できるデータはどれか」がある。このようなデータを発見するためには、データの変動を予測する必要があるが、そのために予測対象データの最近の変動パターンに類似したパターンを検索する方法がある。類似パターン発生後の変化の統計等が予測に有用な材料となることがその理由である。文献[2]では、時系列データに適した類似パターン検索手法を提案し、効果を実証している。しかし、特定の変動が予測できるデータを発見するためには類似パターン検索を何度も繰り返す必要があり、処理が膨大になるとともにユーザに大きな負担がかかるという問題があった。

本稿では、大量の時系列データから効率良く特定の変動が予測できるデータを発見する手法として、時系列データのクラスタリングとパターン検索を融合した2段階解析法を提案する。さらに、同法を搭載した時系列データマイニングシステムを試作し、実験によって2段階解析法の有効性を確認したので報告する。

2 2段階解析法

2.1 2段階解析法の構成

2段階解析法は、大量な過去データの中から近い将来に特定の変動が予測できるデータを効率良く発見する手法である。2段階解析法は、[第1段階]クラスタリングと[第2段階]類似パターン検索から成る（図1）。[第1段階]のクラスタリングによって、全データを特徴ごとにクラスタ分割できる。そして、[第2段階]の類似パターン検索を各クラスタから任意に選んだ代表データの集合から行うため、従来よりも効率良く特定のデータを発見できる。次にクラスタリングと類似パターン検索の手法を示す。

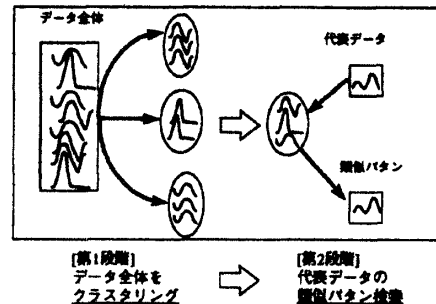


図1: 2段階解析法の構成

2.2 クラスタリング

最も相関係数の高い2つのクラスを統合していくボトムアップ型クラスタリングアルゴリズムの1つである群平均法を用いた。2つの時系列データ $X = \{x_1, x_2, \dots, x_{N_L}\}$ と $Y = \{y_1, y_2, \dots, y_{N_L}\}$ (N_L : 時系列データ長) の相関係数 r は、それぞれの平均を $\bar{x}, \bar{y}$ 、標準偏差を s_x, s_y とするときに、式(1)で定義する。

$$r = \frac{\frac{1}{N_L} \sum_{i=1}^{N_L} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \quad (1)$$

2.3 類似パターン検索

時間軸方向の非線形伸縮を考慮した時系列データのパターン検索を行なうために、ダイナミックプログラミングニューラルネットワーク (DNN) を用いた[2]。DNN は、BP 学習を行なう階層型ニューラルネットワークと DP マッチング[1]を融合したモデルである。ニューラルネット部で、学習済みのパターンと入力パターンの局所的類似度を計算し、DP マッチング部では両者の最適な時間軸の対応付けを決定する。

3 時系列データマイニングシステム

本システムは、(1) 各種時系列データの格納したデータベースが構成するデータベース部、(2) 2段階解析法およびその他の解析を行なう時系列データ解析部、(3) ユーザが対話的に操作し、結果をビジュアルに表示する

*Time Series Data Mining System with 2-Stage Method

†Nobukatsu Kitajima and Tetsuji Tanigawa

‡Human Media Research Laboratories, NEC Corporation

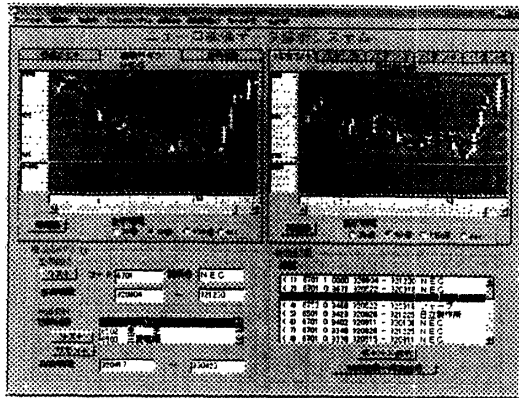


図 2: 時系列データマイニングシステム

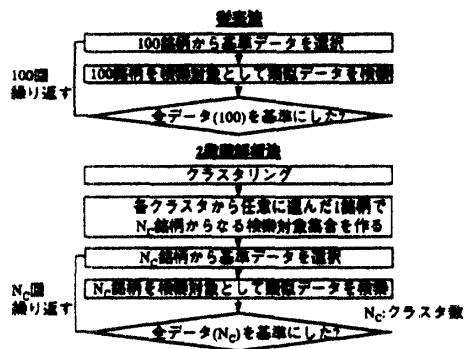


図 3: 従来法および2段階解析法による処理手順

GUI部の3部から成り、Windows95のパーソナルコンピュータで動作する。システムの画面を図2に示す。

4 実験

2段階解析法と従来法の処理効率を比較する実験の結果を示す。

データ 100銘柄の週足株価データ。

実験設定 100銘柄の中から値上がりが予測できる銘柄(有望銘柄)を発見する。

基準期間 最近1/4年のバタンを基準とする類似検索。

クラスタリング 最近1/2年でクラスタリング。

クラス数 $N_c = 10$ 。

実行環境 メモリ:47MB, CPU:Pentium166MHz。

従来法および2段階解析法の処理手順を図3に示す。

類似ボタンを検索する対象期間を0.5年～3年の間で変えた場合の、従来法および2段階解析法による処理時間の比較を表1に、処理時間の比のグラフを図4に示す。

表 1: 従来法と2段階解析法の処理時間比較

(h: 時間, m: 分)			
検索対象期間(年)	0.5	1.0	1.5
従来法	1h12m	3h10m	5h10m
2段階解析法	5m	6m	8m
検索対象期間(年)	2.0	2.5	3.0
従来法	7h10m	9h8m	11h7m
2段階解析法	9m	10m	11m

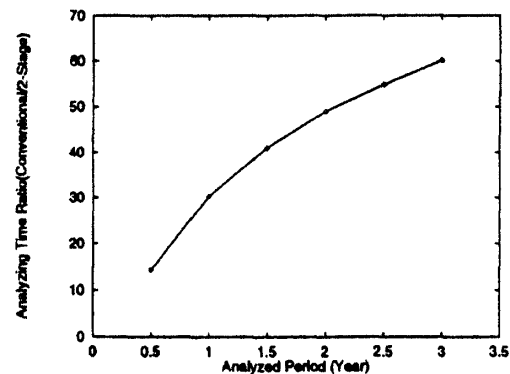


図 4: 従来法と2段階解析法の処理時間の比

図4から2段階解析法を用いることによって、14倍～60倍処理効率が向上することがわかる。両方法による検索結果の質の客観的な比較は困難であるが、2段階解析法は従来法に比べてはるかに効率的であるので、より短い時間で様々な条件での検索が可能となり、結果的に役立つデータをより容易に発見できると考えられる。

5 まとめ

大量の時系列データから効率良く特定のデータを発見する手法である2段階解析法を提案した。さらに同解析法を搭載した時系列データマイニングシステムを試作し、実験によって2段階解析法が従来法よりはるかに効率的であることを確認した。

参考文献

- [1] Sakoe, H. and Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. ASSP-26*, 1, pp. 43-49, 1978.
- [2] Tetsuji Tanigawa and Ken'ichi Kamijo: Stock Price Pattern Matching System - Dynamic Programming Neural Network Approach -, *IJCNN*, Vol.II, pp. 465-471, 1992.6.