2段階解析法を用いた効率的なデータマイニング*

2W－3

北島伸克 谷川哲司†
NEC ヒューマンメディア研究所‡
e-mail: {kitajima, tanigawa}@hml.cl.nec.co.jp

1 はじめに

時系列データマイニングの主要な目的は「予測」である．予測に役立つ情報として，予測対象データの最近の変動パタンに類似したパタンがある [2]．しかし，大量のデータの中から類似パタンを発見しようとする場合，従来の総当たりによる類似パタン検索法を単純に用いると処理時間が膨大になり，ユーザに大きな負担がかかるという問題があった．この問題を解決する方法として，筆者らは2段階解析法を提案し，同法を搭載した時系列データマイニングシステムを構築して，同法による処理時間の削減効果を実証した [1]．ところで，現実に予測をする場面を考えると，単に一定の条件で類似パタン検索を行えば十分なのではなく，様々な条件 (視点) で類似パタンを収集して，それらの視点を総合して予測を行うことが重要である．

本稿では，2段階解析法が処理時間を削減するだけでなく，類似パタンの検索条件を変えることによる視点を変えた予測を支援する効果も併せ持っていることを実験結果とともに示す．

2 2段階解析法

2段階解析法 [1] は，クラスタリングと類似パタン検索を組合わせることによって，大量データから高速に類似パタンを発見する手法である (図 1)．第一段階のクラスタリングには群平均法を用いた [1]．群平均法は，最も相関係数の高い2つのクラスタ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．また，第二段階の類似パタン検索には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) 法 [2] を用いた．DNN 法は，時間軸方向の非線形伸縮を考慮した高精度な時系列データパタン検索を可能とする手法である．

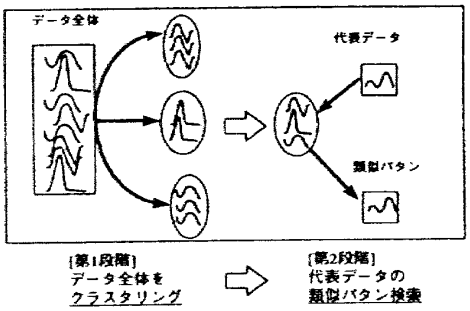

図 1: 2段階解析法の構成

3 2段階解析法におけるクラスタリング

2段階解析法の第一段階であるクラスタリングには2つの効果が考えられる．

第一の効果として処理時間の削減効果がある．その理由は，検索対象データ全体を傾向ごとにクラスタ分割した結果から代表データを選択することによって，検索対象データ空間を絞ることが可能になるからである．この処理時間削減効果は文献 [1] で実証した．

第二の効果としてクラスタリングの条件を変えることが類似パタン検索の視点を変える効果を持つことが考えられる．その理由は，同じ対象データ集合に対してクラスタリングを実行する場合でも，ユーザの選んだ対象期間の傾向によって結果が変化することが予想されるからである．この予想が正しい場合には，特定期間 (特別なイベント後の一定期間等) でクラスタリングすることにより，データ全体を特定期間に依存した傾向によってクラスタ分割できることになる．その結果を用いて代表データを選び，類似パタン検索を行うことによって，視点を変えた類似パタン検索結果を総合した予測が可能になると考えられる．

4 実験

本実験では，100 銘柄の1年分の週足株価データに関して，クラスタ数を固定 (= 10 個) にしてクラスタリン

*Efficient Data Mining Methodology with 2-Stage Method
†Nobukatsu Kitajima and Tetsuji Tanigawa
‡Hunan Media Research Laboratories, NEC Corporation

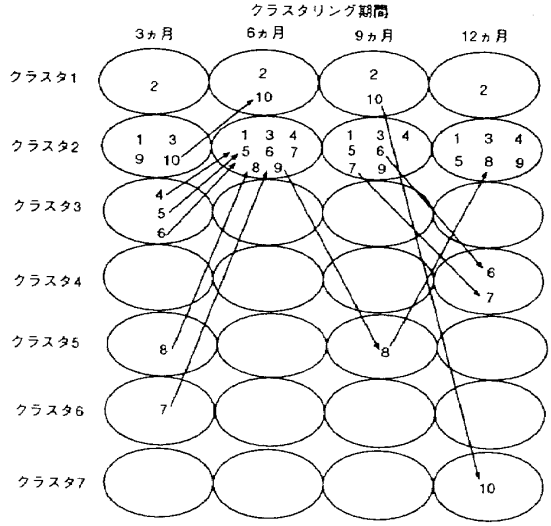

図 2: 異なるクラスタリング期間に対するクラスタ内容の違い (クラスタ内を推移している数字は銘柄番号)

グの際に着目する期間 (以下, クラスタリング期間) を, 3ヶ月, 6ヶ月, 9ヶ月, 12ヶ月と変えた場合のクラスタ内容の違いを調べた. 具体的には 10 銘柄に着目して, それらがどのクラスタに含まれるかを調べた. 着目する 10 銘柄は上記 100 銘柄中に含まれる基準データ (基準期間 : 3ヶ月) に対する総当たり類似パタン検索結果の中の類似度がベスト 10 の銘柄である. 選択した銘柄には類似度順に 1 ～ 10 まで番号を振った. 結果を図 2に示す. 図 2には全 10 個のクラスタの中で上記ベスト 10 の銘柄が含まれている 7 個のクラスタのみを示してあり, クラスタ内の数字は上記ベスト 10 の銘柄の番号である. 例えば, クラスタリング期間を 3ヶ月 → 6ヶ月 → 9ヶ月 → 12ヶ月と変化させた時, 銘柄 1 は常にクラスタ 2 に属しているが, 銘柄 8 は 5 → 2 → 5 → 2 と属するクラスタが変わる. このように同じデータ集合でもクラスタリング期間に依存してクラスタ分割が変化することがわかった. つまり, 同じ 2 つのデータでもクラスタリング期間に依存して, 同じ傾向であると判断できる場合と異なる傾向を持つと判断できる場合があるということである. 例として 6 および 12ヶ月では同じクラスタに属するが, 3 および 9ヶ月では異なるクラスタに属する銘柄 1 と銘柄 8 のパタンを図 3に示す.

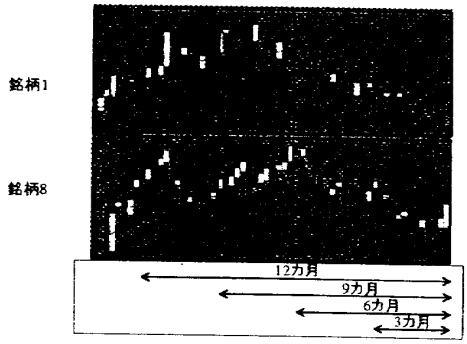

図 3: 期間によって類似性が変わる 2 つのパタン例

5 考察

2 段階解析法の第一段階であるクラスタリングにおいて, クラスタリング期間の選び方によってクラスタリング結果が変化し, 第二段階の類似パタン検索の検索対象選びにも影響を与えることがわかった. 例えば, クラスタリング期間を特別なイベント発生後の一定期間に設定してクラスタリングを実行し, その結果を利用して検索対象データを決めた場合の類似パタン検索結果と, 通常期のクラスタリング結果を利用して検索対象データを決めた場合の類似パタン検索結果とを比較することによって, より高度な予測が可能になると考える.

6 まとめ

2 段階解析法の第一段階であるクラスタリングにおいてクラスタリング期間の選び方によるクラスタ内容の変化を調べた. 実験の結果, クラスタリング期間の選び方によって各期間に依存したクラスタ分割が可能となることがわかった. 2 段階解析法のこの性質を用いることによって, 様々な視点を総合した予測を支援する類似パタン検索が可能になると考える.

参考文献

[1] 北島伸克, 谷川哲司 : 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, 情報処理学会第 55 回全国大会, (2), pp. 543-544, 1997.9.

[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.