

板情報を用いた株式市場の時系列データ分析法

Method of time-series analysis of the stock board

西岡寛兼

Hirokazu NISHIOKA

鳥海不二夫

Fujio TORIUMI

石井健一郎

Kenichiro ISHII

名古屋大学大学院情報科学研究科

Graduate School of Information Science, Nagoya University

Recently the artificial markets are attracting attention of many researchers. To improve the reliability of the artificial market, it is required to show that the artificial market reproduce the real market. In this paper, we propose a method to calculate similarity between behaviors of the markets. In this method, we focus on the board data instead of the price data, and perform time-series analysis. Our simulation shows the effectiveness of the proposed method.

1. 序論

近年、金融市場に対する分析の、新たなアプローチの1つとして期待されているものに、人工市場研究がある。人工市場とは、コンピュータ上に作られた人工的な市場のことである。人工市場研究は、人工市場でのシミュレーション結果を分析することにより、市場に発生した現象のメカニズムを特定したり、市場における新たな事実を発見することを目的としている。しかし、人工市場研究には、シミュレーションと現実との関連付けが十分にされていないという問題がある[和泉 03]。この問題に対して既存の人工市場研究では、実市場で発生した環境を人工市場で再現することで、現実との関連付けを行ってきた。その際、シミュレーション上で起こった価格変動と実市場で起こった価格変動が類似していることを、関連付けの根拠とした。しかし、今日の株価のように異常な値動きを見せる市場を分析する上で、過去の株価変動を再現できたというだけでは、人工市場が実市場と関連付けられている根拠としては不十分である。そのため、人工市場研究の信頼性を高めるために、人工市場と実市場とが十分に類似していることを評価する必要がある。そこで本研究では、市場の動きを観察し、複数の観測データ間の類似性を評価する手法を提案する。

提案手法では、価格ではなく「板」を情報として使用する。板とは、全トレーダーの注文を集計し、表にしたものである。板の例を図1に示す。板には、未約定の注文について、価格別に注文数の合計が表示される。売り注文は左の列、買い注文は右の列に表示されるため、これらは区別できる。

人工市場研究を行う際には、研究者が市場参加者の構成を決定しなければならない。すなわち、市場参加者の構成は、人工市場と実市場を関連付ける際に重要な要素の1つである。また、株式市場において注文とは、資産や取引スタイルなど、トレーダーの持つ様々な要素を反映して出されるものである。すなわち、注文の集計を表示する板には、市場参加者の構成に関する情報が含まれていると考えられる。したがって、人工市場と実市場の類似性を評価する際に、板に注目することは有効であると考えられる。本研究では特に、一度に出される注文の規模を、市場参加者毎に異なる特徴であると考えられる。そして、

売り注文	価格	買い注文
238	10,070	
234	10,060	
441	10,050	
337	10,040	
380	10,030	
312	10,020	
146	10,010	
123	10,000	
	9,990	23
	9,980	535
	9,970	355
	9,960	532
	9,950	337
	9,940	275
	9,930	280
	9,920	361

10,010 円に対する
売り注文の数量 → 146
10,000 円に対する
売り注文の数量 → 123

9,990 円に対する
買い注文の数量 ← 23
9,980 円に対する
買い注文の数量 ← 535

図1 板の例

板から注文の規模に関する情報を抽出し、これを用いて分析を行う。

2. 板情報分析手法

2.1 板情報の形式

提案手法では板を観測して得た板情報を分析対象とする。そこで、予め板情報の形式について述べておく。

板に表示される値は、注文が出される度に变化する。よって、板は時間軸を持つリアルタイムな情報であると言える。したがって、板を観測する際は、例えば「2008年10月1日13時20分40秒の板」というように、板から特定の時点における状態を切り取ることになる。このように、板のある時点における状態を観測したものを、本論文では板情報スナップショットと呼ぶ。

本研究では、時間軸を持つ板から一定間隔で板情報スナップショットを観測することにより、板の標準化を行う。板の標準化によって得られた板情報スナップショットの列を、板情報として扱う。

連絡先: 名古屋大学大学院情報科学研究科

〒464-8603 愛知県名古屋市千種区不老町

E-mail: nishioka@kishii.ss.is.nagoya-u.ac.jp

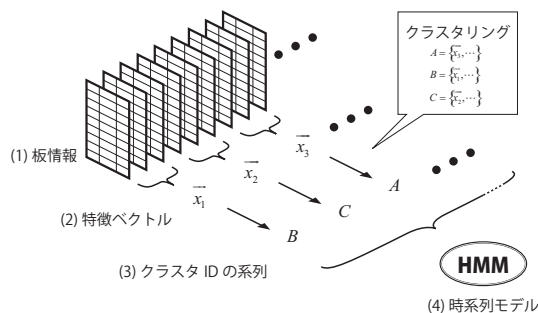


図 2 提案する分析手法の流れ

2.2 分析手法の流れ

本研究で提案する分析手法の流れを図 2 に示す．初めに板情報を，後述する定められた規則に基づいて分割する．そして，分割された板情報の各組から，特徴抽出を行い，特徴ベクトルを生成する．次に，生成された特徴ベクトルに対してクラスタリングを行い，類似した特徴ベクトルを集めたクラスタに分類する．各クラスタに ID を付与することにより，クラスタ ID が時刻順に並べられた系列が生成される．この系列を基に，HMM (Hidden Markov Models) [Rabiner 89] を作成する．HMM を用いることで，その HMM を作成するために使用した学習系列と，任意の入力系列との類似度を計算することが可能となる．

2.3 板情報の分割

板情報を，価格変動の発生直前と発生直後で分割する．なお，本論文において価格変動とは，売り注文価格と買い注文価格の境界が移動することを言う．この分割により生成された，板情報スナップショットの各まとまりを，ユニットと呼ぶ．ユニットは，後述の特徴抽出によって生成される特徴ベクトルと 1 対 1 対応をなす．

2.4 ユニットからの特徴抽出

特徴抽出の流れを図 3 に示す．板情報スナップショットに表示される注文数から，特徴ベクトルを生成する．

まず，前処理として，ユニット内に出現する全ての売り注文数を負の数に変換する．

次に，注文数変化量を求める． i 番目の板情報に表示されている価格 p に対する注文数を $Q(i, p)$ としたとき，注文数変化量 $D(i, p)$ を，

$$D(i, p) = Q(i, p) - Q(i - 1, p) \quad (1)$$

とする．ユニット内に出現する全ての注文数 $Q(i, p)$ に対して求められた注文数変化量 $D(i, p)$ を，そのユニット内で出現する注文数変化量とする．

ここで，注文数変化量の持つ意味について述べる．注文数変化量は，2 つの板情報スナップショットが観測される間に増加した注文数の合計を表す．この合計が，何人のトレーダーによって出された注文の合計であるか，板情報から求めることはできない．しかし，板情報スナップショットを観測する間隔を短くすると，注文数変化量が複数のトレーダーによって出された注文の合計である確率は小さくなる．すなわち，板情報スナップショットを観測する間隔を十分に短くすることによって，注文数変化量の絶対値を，1 人のトレーダーが出した注文の規模として扱うことができる．また，予め行った前処理により，売り注文の増加は負数，買い注文の増加は正数の注文数変化量となり，両者を区別することが可能である．

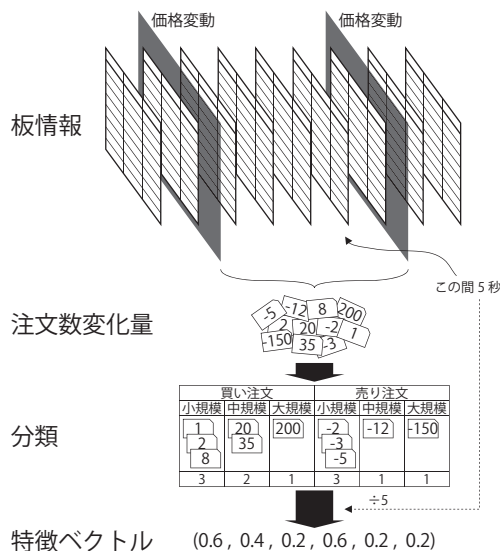


図 3 特徴抽出

最後に，注文数変化量から，特徴ベクトルを生成する．特徴ベクトルの各要素には，「どのような規模の売り注文，あるいは買い注文が何回出されたか」という情報を持たせる．そのために，まず，注文数変化量を，売りか買いかは，前述の通り，注文数変化量の符号が正ならば買い注文，負ならば売り注文として分類する．注文の規模については，注文数変化量の絶対値に境界を定める（例えば，10 未満を小規模，10 以上 100 未満を中規模，100 以上を大規模と定める）ことで，分類する．次に，各分類に属する注文数変化量が，ユニット内に何回出現したかを数える．そして，この出現回数を，このユニットの時間で割ることにより正規化し，注文の分類ごとに並べてベクトルとしたものを特徴ベクトルとする．

2.5 クラスタリング

特徴ベクトルのクラスタリングを行い，類似した特徴ベクトルをまとめたクラスタに分類する．そして，分類された各クラスタにクラスタ ID を付与することにより，ベクトル量子化を行う．この結果，特徴ベクトルを時系列順に並べた列から，クラスタ ID の系列が生成される．

まず，特徴ベクトルの各要素を標準偏差で割ることによって正規化する [石井 98]．次に，正規化した特徴ベクトルのクラスタリングを行う．クラスタリングは，並行 k -means 法を使用する [齋藤 06]．並行 k -means 法は，非階層的クラスタリングを行うアルゴリズムである．

2.6 HMM の作成

クラスタ ID の系列から，クラスタ ID を出力シンボルとみなして HMM を作成する．HMM とは，時系列パターンを確率的に生成するモデルである．HMM の例を図 4 に示す．HMM は，クラスタ ID の系列を学習系列とし，EM アルゴリズムを用いて求める [Levinson 83]．

HMM を作成すると，任意の入力系列が，この HMM によって出力される確率を求めることができる．この出力確率は，入力系列と学習系列との類似性と考えることができる．本研究では，出力確率を入力系列の長さで正規化したものを類似度とする．

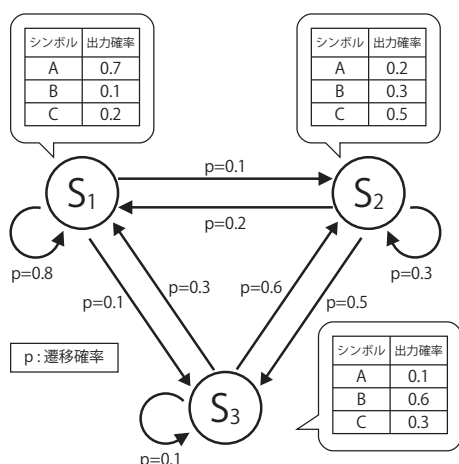


図 4 HMM

3. 提案手法の有効性評価シミュレーション

3.1 目的

本シミュレーションで、提案手法の有効性を確認する。有効性の確認は、2つの観点から行う。

1つは、パラメータに対する頑健性である。提案手法を用いる際には、恣意性のあるパラメータとして以下の3つが存在する。

- クラスタ数 k
- 初期クラスタ
- HMM の状態数

提案手法がこれらのパラメータに大きく影響されることがないことを確認する。

もう1つは、類似性評価能力である。実市場で行われる取引について、1日の内の取引開始直後や取引終了直前における市場の動きは特殊であると言われている。価格変動率や出来高などの指標がこれらの時間帯に特殊な値を示すことが確認されている [渡部 04]。そこで本シミュレーションでは、取引開始直後や取引終了直前と、それ以外の時間帯（平常時間帯）との違いを、提案手法を用いて検出可能か否かを確認する。これにより、提案手法が市場の類似性評価に有効であることを確認する。

3.2 シミュレーション設定

本シミュレーションは、特殊時間帯クラスと平常時間帯クラスの識別問題を定義し、識別率によって評価する。識別は、提案手法により算出される類似度を比較することによって行う。まず、提案手法を用いて特殊時間帯クラス HMM と平常時間帯クラス HMM を作成する。任意の入力系列に対して、両クラスの HMM との類似度をそれぞれ算出し、より高い類似度を示したクラスを識別結果とする。

シミュレーションに使用する板情報は、実市場を観測して収集した。以下の条件で板情報を取得し、45日分のデータを収集した。

- 銘柄は日経 225 先物
- 2008年10月から12月の3ヶ月間
- 板のスナップショットを1秒毎に記録
- 以下の日を除外

表 1 識別問題定義

識別問題 I	
特殊時間帯クラス	前場の取引開始から 30 分間
平常時間帯クラス	10:00 から 10:30 の 30 分間
識別問題 II	
特殊時間帯クラス	後場の取引開始から 30 分間
平常時間帯クラス	13:30 から 14:00 の 30 分間

- サーキットブレーカー発動日
- 12月30日（年末のため市場の稼働が午前のみ）
- 板観測システムの不調により正常な板情報取得ができなかった日

特徴ベクトルを生成する際の注文の規模は、10未満を小規模、10以上100未満を中規模、100以上を大規模とし、買い注文と売り注文合わせて6つに分類した。

本シミュレーションでは、表1に示す2種の識別問題を定義した。いずれの識別問題も、一方の時間帯を特殊時間帯クラス、もう一方の時間帯を平常時間帯クラスとしている。なお、表1において、取引開始とはザラバ方式による取引の開始を意味し、取引終了とはザラバ方式の取引終了を意味する。

学習系列とテスト系列の選択については、44日分のデータを学習系列とし、1日分のデータをテスト系列とする、45重交差検証法を用いた。

3.3 シミュレーション結果と考察

3.3.1 パラメータに対する頑健性

まず、並行 k -means 法のパラメータ k 、及びクラスタの初期設定が、識別率に与える影響を確認した。HMM 状態数を10に固定し、 k を変化させて、識別問題 I について識別率を比較した。 k は5, 10, 20, 50の4種に変化させた。それぞれの k に対して、初期クラスタを変えて、10回の識別を行った。そして、 k の値毎に識別率の平均と標準偏差を求めた。比較結果を表2に示す。

表2より、初期クラスタの変動によるばらつきは小さい。よって、初期クラスタの変動による影響は無視できる範囲であると考えられる。また、 $k \leq 20$ の範囲では、 k の値が与える影響は無視できる範囲であった。しかし、 $k = 50$ の場合は、 $k \leq 20$ の場合と比べて低い識別率となった。これは、 k の値が増えると実際に生成されるクラスタ数も増え、学習系列におけるクラスタ ID 毎の出現回数が減るため、相対的に学習量が不足することが原因である可能性がある。よって、50以上の k が不適切であると結論付けることはできない。

次に、HMM 状態数が識別率に与える影響を確認するシミュレーションを行った。 k を5に固定し、HMM 状態数を変化させて、識別問題 I 及び識別問題 II について識別率を比較した。HMM 状態数は1から15までの15種に変化させた。それぞれの HMM 状態数に対して、初期クラスタを変えて10回の識別を行い、平均を比較した。比較結果を図5に示す。

識別問題 I に関しては、HMM 状態数が1及び2の場合に識別率が低下した。そして、HMM 状態数3以上では識別率に大きな変化は見られない。一方、識別問題 II に関しては、HMM 状態数6以下の場合には識別率に変動が見られる。そして、HMM 状態数9以上では識別率に大きな変化は見られない。以上の結果から、HMM 状態数は9以上の値を用いればよいと考えられる。

表 2 k 毎の識別率比較と乱数による影響

k	5	10	20	50
識別率の平均	0.8633	0.8344	0.8444	0.7789
標準偏差	0.0158	0.0303	0.0314	0.0302

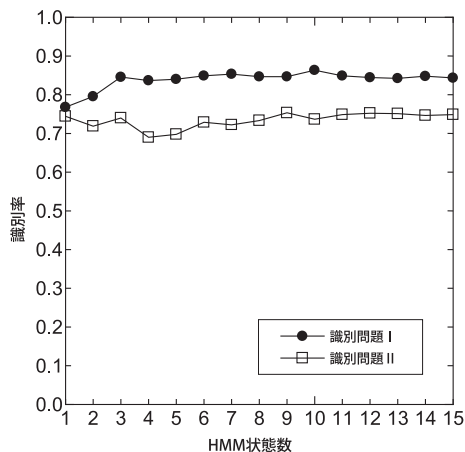


図 5 HMM 状態数-識別率グラフ

3.3.2 類似性評価能力の検証

前節より得られた結果から、 $k = 5$, HMM 状態数 10 とした。なお、評価能力の検証は、並行 k -means 法における初期クラスを変えて、10 回のテストを行って得た識別率の平均によって行った。この結果、識別問題 I で 0.863, 識別問題 II で 0.737 という識別率を得た。この結果から、提案手法によって特殊時間帯と平常時間帯とを識別することが可能であると考えられる。また、朝の取引開始直後、即ち識別問題 I における特殊時間帯クラスの時間帯は、特に市場の動きが特殊であると言われている。識別問題 I では識別問題 II に比べて高い識別率が得られており、提案手法はこの特殊性を評価できたと考えられる。

本シミュレーションで特殊時間帯クラスとして定義した時間帯は、一般に出来高が高いことが確認されている。シミュレーションに使用した系列を観測した 30 分間の出来高を調べ、クラス別に平均を取った結果を表 3 に示す。

また、識別に失敗した系列を観測した 30 分間の出来高を調べ、クラス別に平均を取った結果を表 4 に示す。なお、表 4 は、その系列が本来属しているクラス別の平均を示している。

表 3 では、特殊時間帯クラスの出来高平均の方が高いという結果が確認された。これに対し、識別に失敗した系列のみで出来高平均を求めた表 4 では、表 3 とは逆に平常時間帯クラスの方が高い値を示している。よって、識別に失敗した系列を観測した時間帯は、別の日の同じ時間帯と比べて特殊であったことがわかる。つまり、提案手法はこの特殊性を捉えることに成功したと考えられる。

以上の結果から、提案手法は、市場の類似性評価に有効であると考えられる。

4. 結論

本研究では、人工市場と実市場とがどの程度類似しているかを評価するために、市場から板情報を観測し、これを分析す

表 3 クラス別出来高平均

	特殊時間帯クラス	平常時間帯クラス
識別問題 I	20599.0	9635.7
識別問題 II	16786.0	9246.7

表 4 識別に失敗した系列のクラス別出来高平均

	特殊時間帯クラス	平常時間帯クラス
識別問題 I	7945.8	13791.3
識別問題 II	7823.0	15922.1

ることによって、複数の観測データ間の類似性を評価する手法を提案した。

提案手法の有効性を評価するために、特殊時間帯と平常時間帯との識別問題を 2 種定義し、識別テストを行ったところ、それぞれ約 86%, 約 74% という識別率を得た。また、識別結果を分析した結果、識別に失敗した系列は、出来高について特殊性を備えている系列であることが発見された。これらの結果から、提案手法は、市場の類似性を評価する方法として、有効であることが確認された。

今後の課題として、まず、注文の規模の境界を自動的に、かつ妥当な値に決定する手法の模索が挙げられる。本シミュレーションでは、主観的に境界を決定した。しかし、特に異なる銘柄の取引を比較する場合などにおいては、注文数スケールの差異を吸収できるよう、注文の規模の境界を観測値から自動的に決定する手法が必要であると考えられる。

また、本研究の目的である人工市場と実市場の類似性評価を行う場合においても、提案手法が有効であることを確認する必要がある。

参考文献

- [Levinson 83] Levinson, S., Rabiner, L., and Sondhi, M.: An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, *The Bell System technical journal*, Vol. 62, No. 4, pp. 1035-1074 (1983)
- [Rabiner 89] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286 (1989)
- [石井 98] 石井 健一郎, 上田 修功, 前田 英作, 村瀬 洋: わかりやすいパターン認識, オーム社, 東京 (1998)
- [渡部 04] 渡部 敏明: 東京都立大学 21 世紀 COE プログラム 研究レポート (第 1 回) 「日経 225 先物の価格および取引高の日の変動パターン」 (2004)
- [和泉 03] 和泉 潔: 人工市場: 市場分析の複雑系アプローチ, 森北出版, 東京 (2003)
- [齋藤 06] 齋藤 堯幸, 宿久 洋: 関連性データの解析法 - 多次元尺度構成法とクラスター分析法 -, 共立出版, 東京 (2006)