A-026

推移確率行列を用いた時系列データの予測信頼度
Prediction Reliability using Transition Probability Matrix for Time Series

千葉隆司　松葉育雄
†Ryuji Chiba ‡Ikuo Matsuba

1. まえがき

時系列データの解析手法としては，自己回帰モデルなどといった線形のモデルを用いたものが一般的である．しかし，金融に関係した景気変動や自然界における温度変動などの，一見不規則に見える時系列データに対しては必ずしも満足できる結果が得られるとは限らない．これらの極めて予測困難な事象の背景には，カオスと呼ばれる非線形の力学構造が存在することが予見され，研究されている．また，一般の観測される時系列データにおいては，システムにおけるダイナミクスが動的に変化している．つまり，システムの状態を捉えるのが非常に困難である．そのため，時系列予測値は，システムの状態を捉えることが困難なことと相まって，状況によってその信頼性が大きく異なることが予想される．また，それとは逆に信頼性を予め知ることが出来れば，時系列解析において非常に有用であろう．これは，時系列データの予測を利用している気象予報士や株式ディーラー，経済エコノミストらにとっても望まれていることでもある．

本論文では特に経済時系列データに着目し，次の予測信頼度に関する方法を提案する．初めに変動の相対的変化にのみ注目し文字列化する．次に文字列間の状態遷移をマルコフ過程とみなし，推移確率行列から状態遷移の偏りを抽出する．その偏りが大きいほうが，特徴のある変動と考えられ，予測が容易だと言える．一方，偏りが小さければ，変動が均等に行われるので予測が困難だと言えるだろう．

2. 解析方法

2.1 実験データ

東証株価指数(TOPIX)の１日ごとの終値を用いる．1991/01/04 から 2003/06/24 までの 3075 日分を対象としている．

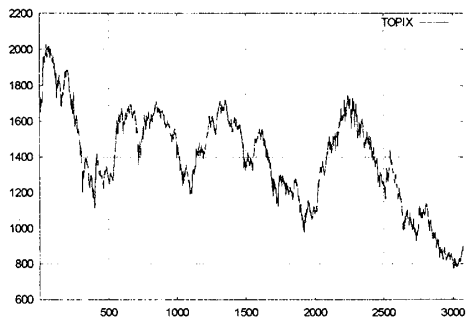

図１ TOPIX の時系列値

† 千葉大学自然科学研究科知能情報工学専攻
‡ 千葉大学工学部情報画像工学科

2.2 文字列化と推移確率行列

経済分野においては，投資した額に対してどれだけ利益が得られるかの割合が重要視されるなど，実際の変動の大小よりも相対的な変化が関心事である場面が多く，特徴量を抽出するにはリターンに変換した後に解析するのが一般的である．ここでは，さらに相対的な上昇下降にのみ着目する．時系列値がその直前の過去よりも大きくなっている場合は up の“u”，小さくなっている場合は down の“d”というように，時系列データをバイナリの文字の配列，すなわち文字列へと変換する(図２)[1].

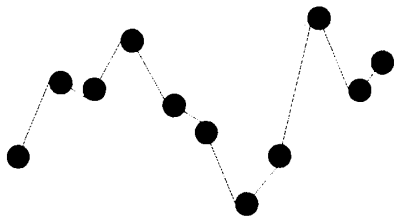

図２ 時系列の文字列化

次にその文字を m 文字ごとに取り出す．これを，m 文字の取りうる全ての組み合わせ(2^m 通り)を状態空間とみなす．図２の例を２文字ごとに取り出すと，“ud”→ “ud” → “dd”→ “uu”→ “du”となる．次に，ある区間の n-d 番目の状態から n 番目の状態へと遷移する確率を求め，推移確率行列 $\boldsymbol{Q}(d)$ として表す．従って，m=2 とし n 番目の確率的状態ベクトル $\boldsymbol{p}(n)$ を導入すると，

$$\boldsymbol{p}(n)=[p_{uu}(n),\ p_{ud}(n),\ p_{du}(n),\ p_{dd}(n)]^T$$

$$\boldsymbol{p}(n)=\boldsymbol{Q}(d)\,\boldsymbol{p}(n-d)$$

と表せる．対象区間全体にわたって，この $\boldsymbol{Q}(d)$ の要素を調べたところ，d>1 ではおおよそ等確率という意味でほとんどランダムとみなせる．このことから，$p(n)$ を決定するプロセスはランダム項 ξ を用いて，

$$\boldsymbol{p}(n)=\boldsymbol{Q}(1)\,\boldsymbol{p}(n-1)+\xi$$

と表せる．故に，TOPIX 時系列は２文字ごとの状態遷移としてとらえた場合に，マルコフ連鎖と仮定することが可能である．また，状態遷移の偏りを推移確率行列 $\boldsymbol{Q}(1)$(以降，単純に $\boldsymbol{Q}$ と表す)を用いて抽出することができる．各状態の遷移において偏りがあるならば予測が困難であり，反対に各状態間の遷移が均質な場合には予測が容易であると考えられる．

例として，TOPIX の 1991/01/04 から 08/09 までの 200 日間

に関して, $Q(1)$を次に示す.

$$\boldsymbol{Q}(1)=\begin{bmatrix}0.28 & 0.24 & 0.37 & 0.27\\ 0.26 & 0.07 & 0.27 & 0.22\\ 0.18 & 0.22 & 0.07 & 0.34\\ 0.28 & 0.46 & 0.29 & 0.17\end{bmatrix}$$

2.3 推移確率行列の固有値・固有ベクトル

推移確率行列 $\boldsymbol{Q}$ は次を満たすことが知られている[2].

① $\boldsymbol{Q}$ は固有値1を有する

② $\boldsymbol{Q}$ の固有値1の重複度は, Q の再帰類の数に等しい

③ $\boldsymbol{Q}$ の固有値の絶対値の最大は1になる

④ Qの固有値1に対する左固有ベクトルの要素は全て1

このことから, 推移確率行列作成対象となる文字列の長さが十分に大きいならば, 非常に稀有な場合を除いて固有値1の重複度は1になる. また, その他の固有値の絶対値は1未満となることから次のことが言える.

$\boldsymbol{Q}$ の右固有ベクトルの1次独立の組 $\phi^{(1)},\ldots,\phi^{(4)}$ と左固有ベクトルの1次独立の双直交な組 $\psi^{(1)},\ldots,\psi^{(4)}$ とを導入する.

$\phi^{(i)}=(\phi_{i1},\ldots,\phi_{i4})$, $\psi^{(i)}=(\psi_{i1},\ldots,\psi_{i4})$ とし, $\bar{\psi}_{jk}$ は ψ_{jk} の共役複素数を表すとしたとき,

$$\Phi=\begin{bmatrix}\phi_{11} & \cdots & \phi_{41}\\ \vdots & & \vdots\\ \phi_{14} & \cdots & \phi_{44}\end{bmatrix},\quad \Psi=\begin{bmatrix}\bar{\psi}_{11} & \cdots & \bar{\psi}_{14}\\ \vdots & & \vdots\\ \bar{\psi}_{41} & \cdots & \bar{\psi}_{44}\end{bmatrix}$$

とする. $\phi^{(1)},\ldots,\phi^{(4)}$ に対応した固有値 $\lambda_1,\ldots,\lambda_4$ を対角成分とした行列 Λ を用いて $\boldsymbol{Q}=\Phi\Lambda\Psi$ のスペクトル表現を有する.

このことから, $\boldsymbol{Q}^m=\Phi\Lambda^m\Psi$ であり, $\boldsymbol{Q}^m$ のi,j成分は

$$(\boldsymbol{Q}^m)_{ij}=\sum_{h=1}^{4}\phi_{hi}\lambda_h^m\bar{\psi}_{hj}$$

であり, $\lambda_1=1$ とし, $m\to\infty$ を取ると $\lim_{m\to\infty}(\boldsymbol{Q}^m)_{ij}=\phi_{1i}$ となる.

以上のことをまとめると, 推移確率行列の固有値1の固有ベクトルの成分の影響は無限時刻後まで残るが, その他の固有ベクトルの成分の影響は0に収束していく. このため, 各状態の遷移の偏りは, 推移確率業悦の固有値1に対する固有ベクトルの成分を調べれば良いと言える.

2.4 固有ベクトルの成分の評価方法

固有値1に対する固有ベクトルの成分の評価方法には, 次の方法を使用した. 成分に似通った値が存在する場合(例として $\phi_{11}=\phi_{12}=\phi_{13}=\phi_{14}=0.25$ といった全ての要素が等しい時や, $\phi_{11}=\phi_{14}=0.32$ で $\phi_{12}=\phi_{13}=0.18$ の2つずつ等しい値が出現する時など)は, 予測が困難であり, 一方で成分の値がばらついている場合(例として $\phi_{11}=0.37$, $\phi_{12}=0.30$, $\phi_{13}=0.20$, $\phi_{14}=0.13$ など)は状態遷移の偏りが大きく, 予測が容易だと考えられる. この直感的な評価を与える, ばらつきの指標 I を提案する. 固有値1の固有ベクトル ϕ_1 の成分を ϕ_{1i} とすると, 指標 I は

$$I=\sum_{i=1}^{4}argmin_{\phi_{1j}}|\phi_{1i}-\phi_{1j}|$$

とする.

3. 解析結果

予測方法は, 線形予測の代表的手法であるARモデルと非線形予測法であるFarmerの方法を用いた. 1991/01/04から2003/06/24までのTOPIXの日次データを500日ごとに6ブロックに分割し, 各ブロックの最初の200日を各モデルにおけるパラメータ推定・推移確率行列作成のための区間とし, 残りの300日分を予測した.

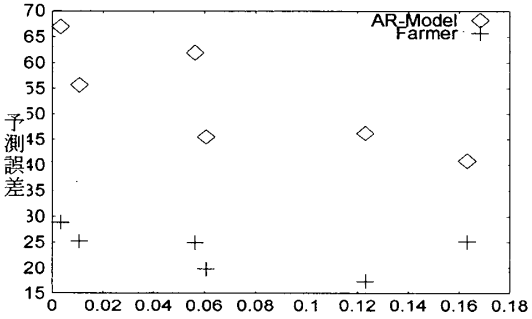

図3 予測誤差とばらつき

ばらつきと予測の平均二乗誤差との関係を図3に示す. ARモデルによる予測では, 概ね負の相関がみられる. 一方, Farmerの方法による予測に関しても, 非常に緩やかであり, 一部例外とみられる場合もあるが, 負の相関を見ることができる.

4. 結論

本研究では, 時系列データをその上昇下降のみに注目した列へと変換し, その連続した変化を一つの状態としてマルコフ連鎖と見なした. そこから, 時系列データの状態遷移の偏りを, 推移確率行列の固有値1の固有ベクトルの成分のばらつきから計算する方法を示した. この偏りの指標が系の予測の信頼度との関係があることを示した. しかしながら, 非線形予測に関しては, パラメータの推定が困難であるという理由などから微妙な結果が得られた. 今後の課題は, 他の時系列データへの適用によるさらなる検証が求められるだろう. また非線形解析方法の成果を待つ一方で, 今回提案した信頼度を逆に予測方法の発展へとつなげることができるかもしれない.

参考文献

[1] N. Vandewalle, M. Ausloos : The n-Zipf analysis of financial data series and biased data series, Physica A 268 (1999) 240-249.

[2] S. Karlin 著, 佐藤健一・佐藤由美子訳 : 確率過程講義, 産業図書.