

ベイジアンネットワークを用いた 株価予測について

左 毅^{†1} 北 栄 輔^{†1}

株価予測のために、過去データを用いた時系列分析に基づくモデルがしばしば用いられる。これらのモデルでは、予測したい株価を過去の株価等と残差の線形和で近似し、分布には正規分布が仮定される。しかし、実データに関するいくつかの知見によれば、株価収益率の頻度分布は必ずしも正規分布しないことが示されている。そうであれば、ホワイトノイズに基づくモデルでは精度良く予測できない可能性がある。そこで、本研究では、ベイジアンネットワークを用いる方法を示す。ところでベイジアンネットワークは離散的な値だけしか扱うことができないので、クラスタリング手法を用いて株価を離散値に変換する。解析例では、日経平均株価をとる。解析結果より、収益率を適切に離散化すれば、一般的な時系列分析モデルよりも精度良く予測できることがわかった。

Stock Prices Prediction by using Bayesian Network

YI ZUO^{†1} and EISUKE KITA^{†1}

Time-series prediction algorithms are very often employed for predicting the stock price. In the algorithms, the stock price is assumed to be the weighted summation of the past stock price and the residual. The distribution is assumed to be according to the normal distribution. However, recent results show that the stock price fluctuation dose not allow the normal distribution. Therefore, in this study, the Bayesian network is employed for the stock price prediction. The Bayesian network can deal with the discrete numbers alone. The use of clustering algorithm transforms the continued number of stock price to the discrete number. NIKKEI stock average is considered as numerical examples. The results show that the prediction accuracy of the present algorithm is better than the time-series prediction algorithms if the stock price is transformed to the discretized number adequately.

1. 緒 論

株価予測には時系列分析に基づく方法が広く用いられている¹⁾。この中には、AR モデル、MA モデル、ARMA モデル、ARCH モデル等がある。これらのモデルでは、予測しようとする株価収益率を説明変数の線形結合で近似し、重み係数を相関解析などにより決定する。このとき、株価収益率の頻度分布は正規乱数に従うと仮定されている。しかし、近年のいくつかの研究によれば、株価変化は正規乱数に完全には従わないことが指摘されている。そこで、本研究では、分布を仮定しないでベイジアンネットワーク²⁾によって株価予測することについて述べる。

ベイジアンネットワークは、確率的なネットワークモデルである。確率変数をノードで表し、変数間の因果関係を非循環有向リンクで表す。さらに、変数間の定量的な依存関係を条件付き確率によって表す。本研究では、このベイジアンネットワークを用いて、過去の株価収益率から今期株価を予測する。ベイジアンネットワークでは株価のような連続値をとる変数を扱うことができないので、クラスタリング手法を用いて株価を離散値に変換する。

解析例として日経平均株価を用いる³⁾。そして、提案手法による予測結果の精度を AR, MA, ARMA, ARCH と比較する。また、ベイジアンネットワークによる予測精度は、株価を離散値に変換する方法にも依存する。そこで、頻度分布を均等に分割して離散化する方法の他に、ウォード法によって離散化する方法を示す。

2. 研究背景

2.1 時系列分析法¹⁾

2.1.1 AR モデル

AR モデルを $AR(p)$ と表記する。 t 期の株価収益率を r_t とすると、これは p 期前までの収益率と、攪乱項 u_t によって次のように表される。

$$r_t = \alpha_0 + \sum_{i=1}^p \alpha_i r_{t-i} + u_t \quad (1)$$

ここで、 α_i はモデルのパラメータである。 u_t は平均 0, 分散 σ^2 のホワイトノイズに従う。

^{†1} 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University

2.1.2 MA モデル

MA モデル $MA(q)$ では、 t 期の株価収益率 r_t を過去の攪乱項によって以下のように表す。

$$r_t = \beta_0 + \sum_{j=1}^q \beta_j u_{t-j} + u_t \quad (2)$$

ここで、 β_j はモデルのパラメータである。

2.1.3 ARMA

ARMA モデル $ARMA(p, q)$ では、 t 期の株価収益率 r_t は株価収益率 r_t と攪乱項 u_t によって次のように表される。

$$r_t = \sum_{i=1}^p \alpha_i r_{t-i} + \sum_{j=1}^q \beta_j u_{t-j} + u_t \quad (3)$$

2.1.4 ARCH

ARCH モデル Engle⁴⁾ $ARCH(p, q)$ では、 t 期の株価収益率 r_t は次のように表される。

$$r_t = \alpha_0 + \sum_{i=1}^p \alpha_i r_{t-i} + u_t \quad (4)$$

ここで、攪乱項 u_t は次式で与えられる。

$$u_t = \sigma_t z_t \quad (5)$$

ここで、 $\sigma_t > 0$ であり、 z_t は平均 0、分散 1 の正規乱数である。

Engle の ARCH モデルによれば、 σ_t^2 は次式で近似される。

$$\sigma_t^2 = \beta_0 + \sum_{j=1}^q \beta_j u_{t-j}^2 \quad (6)$$

2.2 提案手法の概要

上記のモデルでは株価データの分布が正規分布に従うことを仮定している。しかし、実株価の解析から、株価の頻度分布は完全には正規分布に従わないことが指摘されている。原ら⁵⁾ は TOPIX の頻度分布を解析し、ボラティリティを σ とすると、実際の頻度分布ではイベントが $\pm\sigma$ 内にある頻度は正規分布よりも大きく、また、 $\pm 3\sigma$ の外にある頻度は正規分布の 3 倍ほどであることを指摘している。そこで、本研究では正規乱数に基づかず、ベイジアンネットワークを用いて株価収益率の時系列分析を行う。

ベイジアンネットワークを用いれば、確率変数をノードで、変数間の因果関係をリンクで表すことで因果関係を表す非循環有向グラフを得ることができる。本研究では、このベイジアンネットワークを用いて、過去の株価収益率から今期株価収益率を予測する。ただし、ベイジアンネットワークでは確率変数は連続値をとることができないので、クラスタリング手法などを用いて離散化する操作が必要である。

時系列分析手法では、今期の株価収益率を過去の説明変数と攪乱項の線形結合で近似している。これに対して、提案手法では今期株価収益率と過去の株価収益率との間の非線形関係式をベイジアンネットワークによって定義して予測するという見方もできる。

2.3 ベイジアンネットワーク

ベイジアンネットワークでは、確率変数間の定性的な依存関係を確率変数をノード、依存関係をリンクとする非循環有向グラフで可視化し、変数間の定量的な依存関係を条件付き確率によってモデル化する。

確率変数 x_i が x_j に依存していることを $x_j \rightarrow x_i$ と表現し、 x_j を親ノード、 x_i を子ノードと呼ぶ。親ノードが複数ある場合、親ノード集合を $Pa(x_i)$ と表現する。子ノードの親ノードに対する依存関係は条件付き確率 $P(x_i|Pa(x_i))$ で表される。

本研究では、K2Metric^{2),6)} をネットワークの評価値として採用し、K2 アルゴリズム²⁾ を用いてネットワークのグラフ構造を決定する。K2 アルゴリズム²⁾ はよくばり法を元に作られており、全木探索より少ない計算量でネットワークを構築できる。

ベイジアンネットワークの確率推論では、確率変数の確定値から、知りたい確率変数の事後確率を求め、これを用いて期待値などを計算する。

条件付き依存関係が成立しているネットワークにおいて、確率変数の確定値 e に対する知りたい確率変数 x_i の事後確率を $P(x_i|e)$ とする。これは、既に求めた条件付き確率表を用いて周辺化という操作により計算される²⁾。

3. 提案手法

3.1 株価収益率の離散化

解析例として、株価の日次データ³⁾ から求めた株価収益率¹⁾ を扱う。また、ベイジアンネットワークの各ノードが持つ確率変数では連続値を扱うことができない。株価収益率は連続値なので、何らかの方法で離散化する必要がある。離散化した株価の離散値を r^l 、離散値総数を L とすると、離散値の集合は以下のように表される。

$$\{r^1, r^2, \dots, r^L\} \quad (7)$$

離散化のために、本研究では等分割クラスタリングを用いた方法とウォード法を用いた方法を比較する。

クラスタリング法により離散化されたクラスタを C_l 、その重心を c_l とする。離散値 r^l には、各クラスタの重心 c_l をとることとする。つまり、

$$\{r^1, r^2, \dots, r^L\} = \{c_1, c_2, \dots, c_L\} \quad (8)$$

3.1.1 等分割クラスタリングによる離散化

等分割クラスタリングでは、過去の株価収益率の頻度分布図において、それぞれのクラスタに含まれるイベント数がほぼ均等になるように複数のクラスタに等分割する。

3.1.2 ウォード法による離散化

クラスタを C_i 、 C_i の重心を c_i 、 C_i に含まれたデータを z とする。ウォード法では、各対象からその対象を含むクラスタの重心までの距離の二乗の総和を最小化する。ウォード法の評価式は次式で表される。

$$D(C_i, C_j) = E(C_i \cup C_j) - E(C_i) - E(C_j) \quad (9)$$

ただし

$$E(C_i) = \sum_{z \in C_i} d(z, c_i)^2 \quad (10)$$

ここで、 $d(z, c_i)$ は z と c_i のユークリッド距離を示す。

3.2 株価収益率の推定

クラスタリングにより求めた過去の株価収益率の離散値からネットワークを決定する。

本研究では K2 アルゴリズム²⁾ を用いているので、あらかじめ変数間の全順序関係がわかっている必要がある。株価では時系列に基づく全順序関係があるので、これに従ってネットワークを決定する。

離散化された過去の株価収益率からネットワークを決定し、これを B とする。これを用いて株価収益率 $r_t = r^l$ となる確率を推定した結果を $P(r^l|B)$ と定義する。 t 期の株価収益率として、 $P(r^l|B)$ が最大となる r^l を選択する。つまり、

$$r_t = \arg \max_{r^l} (P(r^l|B)) \quad (11)$$

3.3 アルゴリズム

アルゴリズムを整理し直すと以下のようになる。

- (1) 株価収益率を離散化する。
- (2) 離散化された過去の株価収益率によりベイジアンネットワーク B を決定する。

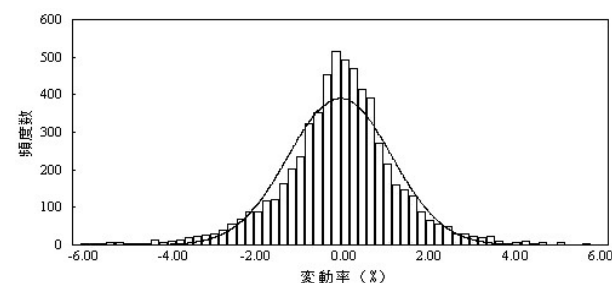


図 1 日経平均株価の頻度分布図

Fig.1 Histogram of NIKKEI Stock Average

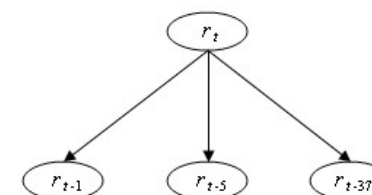


図 2 日経平均株価から決定されたネットワーク (等分割クラスタリング)

Fig.2 Bayesian network determined from NIKKEI225 (Uniform clustering)

- (3) B を用いて式 (11) より株価収益率を予測する。

4. 数値実験

例題として日経平均株価の 1 日次収益率の頻度分布を図 1 に示す。1985 年 2 月 22 日から 2008 年 12 月 30 日までの 6000 日間の日次株価収益率からベイジアンネットワークを決定する。それを 2008 年 12 月 1 日から 2008 年 12 月 30 日の株価を予測に適用し、相関係数 (CC) と平均二乗誤差 (RMSE) により実際の株価と比較する。

4.1 等分割クラスタリングによるネットワーク

クラスタ数を 6 個として、3.1.1 節で説明した等分割クラスタリングで離散化して予測を行う。データから決定されたベイジアンネットワークを図 2 に示す。この場合、株価 r_t は 1 日前の株価 r_{t-1} 、5 日前の株価 r_{t-5} 、37 日前の株価 r_{t-37} と関連づけられていることがわかる。

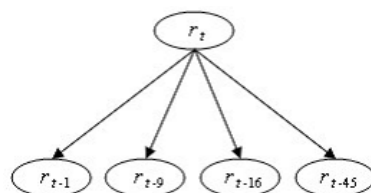


図 3 日経平均株価から決定されたネットワーク (ウォード法)
Fig. 3 Bayesian network determined from NIKKEI225 (Ward method)

表 1 予測値と実測値の比較

Table 1 Comparison of predicted and actual stock prices

	最大誤差	最小誤差	CC	RMSE
BN	5.5203	0.4172	0.7785	2.7521
BN2	5.4433	0.0254	0.7940	2.3205
AR(2)	6.4808	0.0056	0.6928	2.7452
MA(2)	6.4808	0.0399	0.6942	2.7345
ARMA(2,2)	6.6313	0.2826	0.6840	2.7751
ARCH(2,9)	6.5119	0.1069	0.6974	2.7331

4.2 ウォード法によるネットワーク

クラスタ数を 6 個として、3.1.2 節で説明したウォード法で離散化して予測を行う。この離散化を用いて決定したネットワークを図 3 に示す。この場合、株価 r_t は 1 日前の株価 r_{t-1} 、9 日前の株価 r_{t-9} 、16 日前の株価 r_{t-16} 、45 日前の株価 r_{t-45} と関連づけられており、等分割クラスタリングの場合と異なっていることがわかる。

4.3 精度比較

予測結果を実際の株価と比較した結果を表 1 に示す。ここで、AR, MA, ARMA, ARCH は、それぞれ AR, MA, ARMA, ARCH による結果を示す。また、BN と BN2 は等分割クラスタリングとウォード法によって離散化したデータを用いた結果を示す。それぞれにおいて、1000 回シミュレーションを行い、その予測平均値を示す。

等分割クラスタリングによる結果 BN は、AR, MA, ARMA, ARCH モデルより相関係数は高く、平均二乗誤差はほとんど同じであることがわかる。さらに、ウォード法による結果 BN2 では、BN2 の予測結果の相関係数は等分割クラスタリングや ARCH モデルよりも高くなっている。また、平均二乗誤差についてみると、BN2 の誤差は ARCH モデルなどよりも小さくなっている。このことから、離散化法を変更することで、提案手法の精度が

改善することがわかる。

5. 結 論

本研究では、ベイジアンネットワークを用いて株価収益率を予測する方法について述べた。提案手法では、株価収益率の分布をクラスタリング法により離散化し、離散値を用いてベイジアンネットワークを決定する。決定したネットワークを用いて今期の株価収益率を予測する。数値シミュレーションには日経平均株価をとり、提案手法による予測結果を AR モデル、MA モデル、ARMA モデル、ARCH モデルと比較した。

実験では日経平均株価を用いた。株価収益率の離散化法として等分割クラスタリングとウォード法を用いて比較した。等分割クラスタリングを用いる方法では、AR モデル、MA モデル、ARMA モデル、ARCH モデルに比べて、提案手法は同等の精度を示し、相関係数では良い結果を示した。これに対して、ウォード法を用いる方法と提案法の誤差は 15% ほど小さくなり、相関係数も改善した。

参 考 文 献

- 1) 渡部敏明. 日本の株式市場におけるボラティリティの変化. 三菱経済研究所, 1997.
- 2) 繁榊算男, 本村陽一, 植野真臣. ベイジアンネットワーク概説. 培風館, 2006.
- 3) Yahoo!ファイナンス. <http://quote.yahoo.co.jp/>.
- 4) R.F. Engle and VK. Ng. Measuring and testing the impact of news on volatility. *Journal of Finance*, Vol.48, No.5, pp. 1749-78, 1993.
- 5) 原章, 長尾智晴. 自動グループ構成手法 ADG を用いた人工株式市場の構築. 情報処理学会論文誌, Vol.41, No.4, pp. 1063-1072, 2000.
- 6) D.Heckerman, D.Geiger, and D.Chickering. Learning bayesian networks:the combination of knowledge and statistical data. *Machine Learning*, Vol.20, pp. 197-243, 1995.