

ベイジアンネットワークを用いた 株価予測法の精度改善

左 毅^{†1} 原田 昌 朗^{†1} 北 栄 輔^{†1}

これまでに、著者らのグループではベイジアンネットワークを用いた株価収益率の予測法を提案し、日経平均株価とトヨタ自動車株価を例として時系列分析法と比較した結果、時系列分析法よりも高い精度を示した。本研究では、従来法の精度を向上する手法について述べる。従来法では、株価収益率をノードとしてベイジアンネットワークを構築した。これに対して、改善手法では、株価収益率ノードに加えて株価収益率の予測精度ノードを用いてベイジアンネットワークを構築し、株価収益率を予測する。予測精度において比較した結果、提案手法は、時系列分析法に比べて30%、ベイジアンネットワークを用いた従来法に比べて10%程度精度向上できることがわかった。

Improvement of Stock Price Forecast using Bayesian Network

YI ZUO,^{†1} MASA AKI HARADA^{†1} and EISUKE KITA^{†1}

Authors have presented the forecast scheme of the stock price earning ration (PER) by using Bayesian network. The forecast scheme was compared with the time-series forecast schemes, which are popular traditional ones, in the PER prediction of NIKKEI stock average and Toyota motor corporation stock price. The present scheme showed more accurate forecast results than the time-series schemes. In this study, we will present the improved scheme of the PER forecast scheme using Bayesian network. In the previous scheme, Bayesian network for PER forecast was constructed from the PER nodes alone. The improved scheme adopts as the nodes, in addition to the PER node, the PER forecast error node. Numerical results show that the forecast accuracy of the present scheme is improved by 30% against the time-series forecast schemes and by 10% against the previous scheme using Bayesian network.

1. 緒 論

株価予測の代表的な手法は時系列分析である¹⁾。このなかには、Auto Regressive (AR) モデル、Moving Average (MA) モデル、AutoRegressive Moving Average (ARMA) モデル、AutoRegressive Conditional Heteroskedasticity (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り、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ので、正規乱数を用いたモデルでは、十分な予測精度を得られない可能性がある。そこで、著者らは、以前の研究においてベイジアンネットワーク²⁾を用いた株価収益率推定法を提案した³⁾。本研究では、以前に提案した手法の予測精度改善方法について検討する。提案手法では、従来の手法で過去の株価収益率を予測した後、その予測値の誤差を評価する。誤差分布をクラスタリング法によって離散化して誤差分布の離散値を求め、誤差ノードを定義する。誤差ノードを株価収益率ノードと共に用いてベイジアンネットワークを構築し、これを株価収益率予測に用いる。また、従来の方法では、求められたネットワークはNaive ベイズ構造であった。そこで、本研究では、計算コストを下げるためにNaive ベイズ構造を仮定してネットワーク構造を探索する。解析例として日経平均株価⁴⁾をとり、提案手法による予測結果を時系列分析法や従来の方法³⁾と比較する。

2. ネットワーク探索アルゴリズム

2.1 Naive ベイズ構造と条件付き確率

確率変数 x_0 が x_i に依存していることを $x_i \rightarrow x_0$ と表現し、 x_i を親ノード、 x_0 を子ノードと呼ぶ。 x_0 の親ノードが複数ある場合、その親ノード集合を $Pa(x_0)$ と表現することにする。親ノード集合に属するノードを x_i 、その総数を M とすると $Pa(x_0)$ は次式のように定義される。

$$Pa(x_0) = \{x_1, x_2, \dots, x_M\} \quad (1)$$

子ノードの親ノード集合に対する依存関係は条件付き確率 $P(x_0|Pa(x_0))$ で表される。これは次式で与えられる。

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

$$P(x_0|Pa(x_0)) = \frac{P(x_0)P(Pa(x_0)|x_0)}{P(Pa(x_0))} = \frac{P(x_0) \prod_{i=1}^M P(x_i|x_0)}{P(Pa(x_0))} \quad (2)$$

2.2 K2Metric

本研究では、K2Metric^(2),5) をネットワークの評価値として採用する。全ノード数を N 、子ノード x_0 が取りうる状態の総数を L 、親ノード集合 $Pa(x_0)$ が取りうる状態の総数を M と表す。また、親ノード集合 $Pa(x_0)$ が状態 Y^j をノード x_0 が状態 X^k を取る場合の個数を N_{jk} とする。事前分布が一様分布であるとすると、与えられたノード集合から構築されたネットワークについて、K2Metric^(2),5),6) は次式で与えられる。

$$K2 = \prod_{j=1}^M \frac{(L-1)!}{(N_j + L - 1)!} \prod_{k=1}^L N_{jk}! \quad (3)$$

$$N_j = \sum_{k=1}^L N_{jk} \quad (4)$$

2.3 グラフ構造の探索

ネットワークのグラフ構造を決定するために、K2 アルゴリズム⁽²⁾ を Naive ベイズ構造に対して修正したアルゴリズムを用いる。子ノードを x_0 とし、親ノードとなりうるノード x_i の総数を N とするときグラフ構造を探索するアルゴリズムは以下になる。

- (1) x_0 に対する親ノード集合 $Pa(x_0)$ を空集合 ϕ として定義する。
- (2) x_0 と $Pa(x_0)$ からなるネットワークの K2Metric を S_{best} とする。
- (3) $i = 1$ とする。
- (4) x_i を $Pa(x_0)$ に加える。
- (5) x_0 と $Pa(x_0)$ からなるネットワークの K2Metric を S とする。
- (6) $S \leq S_{best}$ ならば、 x_i を $Pa(x_0)$ から除外する。
- (7) $S > S_{best}$ ならば、 $S = S_{best}$ とする。
- (8) $i = i + 1$ とし、 $i \leq N$ ならば、ステップ 4 へ戻る。
- (9) x_0 と $Pa(x_0)$ からなるネットワークを B とする。

3. 従来手法

3.1 株価収益率の離散化

本研究では、次式で定義される株価収益率⁽¹⁾ を扱う。

$$r_t = (\ln P_t - \ln P_{t-1}) \times 100 \quad (5)$$

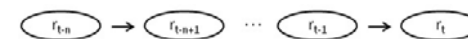


図 1 株価収益率の全順序関係
Fig.1 Total order of stock volatility

ここで、 P_t は t 期の株価終値を、 r_t は t 期の株価収益率を示す。これは、株価指数の終値の日次データ⁽⁴⁾ から求める。

株価収益率は連続値だが、ベイジアンネットワークで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。そこで、何らかの方法によって株価収益率分布を離散化して離散値を得る。以前の研究⁽³⁾ では、均等クラスタリング法とウォード法によるクラスタリングを比較し、ウォード法のほうが良い結果を示した。(ウォード法については、4.2 節で説明する。)

株価収益率離散値を 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。

$$\{r^1, r^2, \dots, r^L\} \quad (6)$$

3.2 ネットワークの構成と株価収益率の推定

クラスタリングにより離散化された過去の株価収益率をベイジアンネットワークに適用し、予測用のネットワークを決定する。ネットワークを決定するために、本研究では K2 アルゴリズム⁽²⁾ を用いる。そこで、株価の時系列に基づく全順序関係に従って、ネットワークを決定する(図 1)。

離散化された過去の株価収益率から決定されたネットワークを B 、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、

$$r_t = \arg \max_{r^l} (P(r^l|B)) \quad (7)$$

3.3 従来手法のアルゴリズム

アルゴリズムを整理し直すと以下になる。

- (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をノードとしてベイジアンネットワーク B を決定する。
- (3) B を用いて式 (7) より株価収益率を予測する。

4. 提案手法

4.1 提案手法のアルゴリズム

最初に提案手法のアルゴリズムについて示し、従来手法との違いについて述べる。



図2 株価収益率と予測誤差の全順序関係
 Fig.2 Total order of PER and prediction error

- (1) 3節で述べた従来手法により、すでに測定された過去の株価収益率を予測する。
- (2) 過去の株価収益率について、予測値と実測値から誤差を求める。
- (3) 誤差分布をウォード法によって離散化し、誤差の離散値を求める。
- (4) 過去の株価収益率と、それに対する評価から求めた誤差の離散値をノードとしてベイジアンネットワークを構築し、これを B' とする。
- (5) B' を用いて最尤推定して求めた株価収益率を予測値として採用する。

提案手法は、従来法に誤差を評価して離散化する過程を追加した手法となっている。そこで、以下では、誤差を評価して離散化する過程とネットワーク構築に関わる過程について詳しく述べることにする。

4.2 ウォード法

離散化のために、本研究ではウォード法を用いる。クラスタを C_i 、 C_i の重心を c_i 、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。

$$D(C_i, C_j) = E(C_i \cup C_j) - E(C_i) - E(C_j) \quad (8)$$

ただし

$$E(C_i) = \sum_{z \in C_i} d(z, c_i)^2 \quad (9)$$

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

4.3 予測誤差の離散化

すでに測定された過去の t 期において、式 (7) より求めた株価収益率の予測値を r_t 、実測値を \bar{r}_t とする。このとき、予測誤差 u_t は次式で与えられる。

$$u_t = r_t - \bar{r}_t \quad (10)$$

予測誤差は連続値なので、ベイジアンネットワークで適用できるように4.2節で示したウォード法によって離散化する。

予測誤差離散値を u^l 、離散値総数を L とすると、予測誤差離散値の集合は以下のように表される。

$$\{u^1, u^2, \dots, u^L\} \quad (11)$$

4.4 ネットワークの構成と株価収益率の推定

過去の株価収益率離散値とその予測誤差の離散値を用いて、図2に示す全順序関係に従って、2節のアルゴリズムによりネットワークを決定する。

決定されたネットワークを B' 、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 $P(r^l|B')$ とする。 $P(r^l|B')$ が最大となる r^l を t 期の株価収益率とする。つまり、

$$r_t = \arg \max_{r^l} (P(r^l|B')) \quad (12)$$

5. 数値実験

例題として日経平均株価をとる。ベイジアンネットワークを決定するために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を使用する。そして、2008年12月1日から2008年12月30日の株価収益率を予測する。そして、本研究で提案する手法を時系列分析手法や従来の研究で提案した方法³⁾と比較する。既に述べたように、時系列分析手法のモデルパラメータ次数 p, q は、これらを1から10までとってシミュレーションを行い、赤池情報量規準 AIC(式(13))が最も小さい値を示す場合のモデルパラメータを用いる。

$$AIC = \ln \hat{\sigma}^2 + \frac{2(p+q)}{T} \quad (13)$$

ここで、 $\hat{\sigma}$ はモデルの残差 $\epsilon_1, \epsilon_2, \dots, \epsilon_T$ の標本分散であり、 T は残差の総数である。

ベイジアンネットワークを用いた手法のうち従来法と提案手法についても、1期前から10期前までの株価収益率や予測誤差のノードを用いてシミュレーションを行い、AIC(式(13))が最も小さい値を示すベイジアンネットワークを用いる。このとき、先の研究³⁾を参考にし、株価変化率と誤差の両方について、離散値総数 $L = 6$ とする。

3節のアルゴリズムにより、株価収益率のみで決定されたネットワークを図3に示す³⁾。この場合、株価 r_t は1日前の株価 r_{t-1} 、4日前の株価 r_{t-4} 、9日前の株価 r_{t-9} と関連づけられていることがわかる。次に、4節のアルゴリズムにより、株価収益率と予測誤差で決定されたネットワークを図4に示す。この場合、株価 r_t は1日前の株価 r_{t-1} 、4日前の株価 r_{t-4} 、9日前の株価 r_{t-9} 、3日前の誤差 u_{t-3} 、4日前の誤差 u_{t-4} と関連づけられていることがわかる。予測誤差ノードを取り入れると、異なるネットワークが導かれることがわかる。

求めたネットワークで予測を行った結果を、時系列分析法や従来法と比較した結果を図5に示す。ここで、AR, MA, ARMA, ARCH は、それぞれの時系列分析法による結果を、

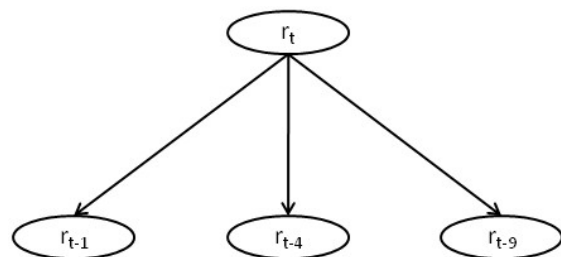


図3 株価収益率のみで決定されたベイジアンネットワーク (日経平均)
Fig.3 Bayesian network with NIKKEI stock average PER only

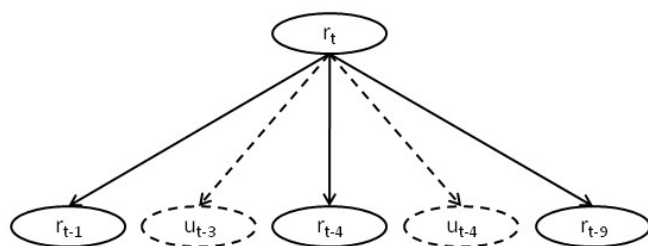


図4 株価収益率と誤差で決定されたベイジアンネットワーク (日経平均)
Fig.4 Bayesian network with Nikkei stock average PER and Error

BN(Previous) は図3のネットワークによる予測結果を, BN(Present) は図4のネットワークによる予測結果を示す. それぞれの数値は相関係数(CC), 平均二乗誤差(RMSE), 最大誤差(MaxE), 最小誤差(MinE)を示す.

図5から, 提案手法は最も大きな相関係数を示しており, 平均二乗誤差と最大誤差も最小となっていることがわかる. 最小誤差はどの方法でも同程度であるが, 平均二乗誤差は提案手法では時系列分析法よりも30%程度, 従来手法BN(Present)よりも10%程度改善されていることがわかる.

6. 結論

本研究では, 著者らがこれまでに提案しているベイジアンネットワークを用いた株価収益率の予測法の精度改善手法について述べた.

従来法では, 過去の株価収益率が特定の値を取るときに, 次の時点の株価収益率が特定の

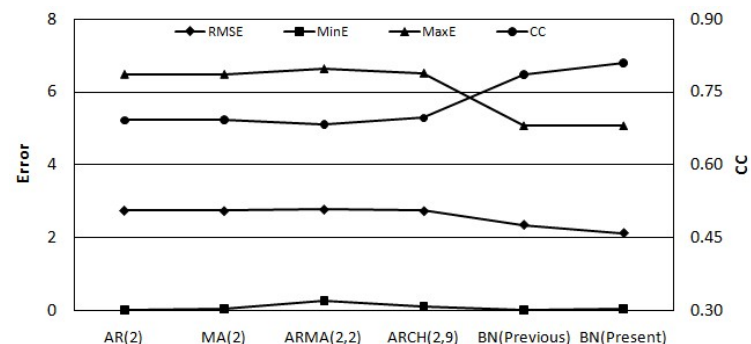


図5 予測値と実測値の比較 (日経平均株価)
Fig.5 Comparison of predicted and actual stock prices (Nikkei stock average)

値を取る確率を推測するためのベイジアンネットワークを構築し, これを将来株価の予測に用いる. これに対して, 提案した方法では, 最初に従来法で過去の既知の株価を予測し, 予測精度を評価する. この予測精度を株価収益率とともに用いてノードを定義してベイジアンネットワークを構築する.

日経平均の株価収益率に適用した結果, 平均誤差では, 提案手法は時系列分析法よりも30%程度, バイジアンネットワークを用いた従来手法よりも10%程度改善された.

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) 繁樹算男, 本村陽一, 植野真臣. バイジアンネットワーク概説. 培風館, 2006.
- 3) 左毅, 北栄輔. バイジアンネットワークを用いた株価予測について. 情報処理学会論文誌トランザクション 数理モデル化と応用, Vol.3, No.3, pp. 80-90, 2010.
- 4) Yahoo!ファイナンス. <http://quote.yahoo.co.jp/>.
- 5) 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.
- 6) G.F. Cooper and Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, Vol.9, No.4, pp. 309-347, 1992.