Expert Systems with Applications 39 (2012) 6729–6737

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Stock price forecast using Bayesian network

Yi Zuo, Eisuke Kita *

Nagoya University, Nagoya 464-8601, Japan

ARTICLE INFO

Keywords:
Stock price
Bayesian network
Uniform clustering
Ward method

ABSTRACT

Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. This paper describes the price earnings ratio (P/E ratio) forecast by using Bayesian network. Firstly, the use of clustering algorithm transforms the continuous P/E ratio to the set of digitized values. The Bayesian network for the P/E ratio forecast is determined from the set of the digitized values. NIKKEI stock average (NIKKEI225) and Toyota motor corporation stock price are considered as numerical examples. The results show that the forecast accuracy of the present algorithm is better than that of the traditional time-series forecast algorithms in comparison of their correlation coefficient and the root mean square error.

1. Introduction

Stock price forecast is very important for planning of business activity and the national economy. Several time-series forecast algorithms have been applied successively for the stock price forecast (Box, Jenkins, & Reinsel, 1994; Brockwell & Davis, 2009). Auto Regressive (AR) model, Moving Average (MA) model, Auto Regressive Moving Average (ARMA) model and AutoRegressive Conditional Heteroskedasticity (ARCH) model (Engle & Ng, 1993) are very popular algorithms. AR model approximates the stock price with previous stock prices and MA model uses, instead of the previous stock prices, the previous error terms. ARMA model is the combinational model of AR and MA models. In ARCH model, the stock price is approximated with the linear combination of the previous stock prices and the error term. The volatility of the error term is approximated with the previous error terms. ARCH model was presented by Engle and Ng (1993) in 1980s. After that, many researchers have presented several improved models from ARCH such as Generalized AutoregRessive Conditional Heteroskedasticity (GARCH) model (Bollerslevb, 1986), Exponential General AutoRegressive Conditional Heteroskedastic (EGARCH) model (Nelson, 1991) and so on.

The time-series forecast algorithms usually represent the error distribution according to the normal distribution. Recent studies point out that the distribution of the stock price fluctuation does not follow the normal distribution (Takayasu, 2006). Especially, the analysis of actual stock data reveal that the deviation around $\pm\sigma$ and $\pm 3\sigma$ is thicker than the normal distribution. The algorithm based on the normal distribution may not forecast the stock price accurately. Therefore, the stock price forecast by using Bayesian network (Ben-Gal, 2007; Pearl & Russell, 2002) is presented in this study.

Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. The use of the Bayesian network enables the stock price forecast without white noise model. Although the stock price is continuous value, the Bayesian network can deal with the discrete (digitized) values alone. The stock price distribution is digitized firstly by using the clustering algorithms and then, the Bayesian network is used for modeling the stochastic dependencies among the digitized values of the previous stock price.

NIKKEI stock average and Toyota motor corporation stock price are considered as examples. While the P/E ratio distribution of NIKKEI stock average is relatively similar to the normal distribution, the P/E ratio distribution of Toyota motor corporation stock price is far from the normal distribution. The present method is compared with AR, MA, ARMA and ARCH on their forecast accuracy. Since the present method depends on the clustering algorithms, two clustering algorithms; uniform clustering and the Ward method, are compared and the effectiveness of the number of clusters (set of digitized numbers) is also discussed.

The remaining part of the manuscript is as follows. In Section 2, time-series forecast algorithms are explained briefly. Bayesian network algorithm is described in Section 3 and the present algorithm is explained in Section 4. Numerical results are shown in Section 5. The results are summarized again in Section 6.

* Corresponding author.
E-mail address: kita@is.nagoya-u.ac.jp (E. Kita).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.12.035

2. Background

2.1. Time-series forecast algorithms (Box et al., 1994; Brockwell & Davis, 2009)

2.1.1. AR Model

The notation r_t denotes the price earnings ratio (P/E ratio) of the stock at time t. In AR model AR(p), the P/E ratio r_t is approximated with the previous P/E ratio r_{t-i} ($i = 1,\ldots,p$) and the error term u_t as follows:

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

where α_i ($i = 0,\ldots,p$) is the model parameter. The error term u_t is a random variable from the normal distribution centered at 0 with standard deviation equal to σ^2.

2.1.2. MA model

In MA model MA(q), the P/E ratio r_t is approximated with the previous error term u_{t-j} ($j = 1,\ldots,q$) as follows:

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

where β_j ($j = 0,\ldots,q$) is the model parameter.

2.1.3. ARMA model

ARMA model is the combinational model of AR and MA models. In ARMA model ARMA(p,q), the P/E ratio r_t is approximated as follows:

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

2.1.4. ARCH model

In ARCH model ARCH(p,q), the P/E ratio r_t at time t is approximated as follows:

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

The error term u_t is given as follows:

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

where $\sigma_t > 0$ and the function z_t is a random variable from the normal distribution centered at 0 with standard deviation equal to 1.

The volatility σ_t^2 is approximated with

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

2.1.5. Determination of model parameters

In each model, the model parameters p and q are taken from $p = 0,1,\ldots,10$ and $q = 0,1,\ldots,10$. Akaike's Information Criterion (AIC) is estimated in all cases. The parameters p and q for maximum AIC are adopted.

The AIC is given as follows:

$$\mathrm{AIC} = \ln \hat{\sigma}^2 + \frac{2(p+q)}{T} \quad (7)$$

where $\hat{\sigma}$ is the volatility estimated from the model error $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$.

2.2. Concept of present method

In the time-series forecast algorithms, it is assumed that the stock price fluctuation allows the normal distribution. Recent studies, however, point out that the distribution of the stock price fluctuation does not follow the normal distribution (Takayasu, 2006). Fig. 1 shows the P/E ratio of NIKKEI stock average. This figure is plotted with the date as the horizontal axis and the P/E ratio as the vertical axis. The frequency distribution of the P/E ratio as well as the normal distribution is shown in Fig. 2. This figure is plotted with the P/E ratio as the horizontal axis and the frequency distribution as the vertical axis. Fig. 2 shows that the frequency distribution of the NIKKEI stock average P/E ratio is a little far from the normal distribution. Fig. 3 illustrates the P/E ratio frequency distribution of Toyota motor corporation stock price as well as the normal distribution. It is clearer in this figure that the frequency distribution is very far from the normal distribution and especially, the deviation around $\pm\sigma$ and $\pm 3\sigma$ is conspicuous. As a result, the normal distribution may not forecast the stock price accurately. For overcoming this difficulty, Bayesian network is applied for the stock price forecast in this study.

The Bayesian network represents a set of random variables and their conditional dependencies via a directed acyclic graph. The Bayesian network model can forecast the error terms without white noise model. Although the stock price is continuous value, the Bayesian network can deal with the discrete values alone. Therefore, the stock price distribution is digitized firstly by using the clustering algorithms.

In the time-series forecast algorithm, the P/E ratio is approximated with the linear combination of the previous P/E ratio values and the error term. In the present method, the P/E ratio is modeled in the non-linear model with Bayesian network Fig. 4 shows the relationship between the AR model, one of the ordinary models, and the Bayesian network.

3. Bayesian network

3.1. Conditional probability table

In Bayesian network, the conditional dependencies among a set of random variables are represented with a directed acyclic graph.

If the random variable x_i depends on the random variable x_j, the variable x_j and x_i are called as a parent and a child, respectively. Their dependency is represented with $x_j \to x_i$. If more than one parents exist for the child x_i, the notation $Pa(x_i)$ denotes the parents set for x_i. Conditional dependency probability between x_j and x_i is represented with $P(x_i|x_j)$, which means the conditional probability of x_i given x_j.

The strength of relationships between random variables is quantified with the conditional probability table (CPT). The notation Y^m and X^l denote the mth state of $Pa(x_i)$ and the lth state of x_i, respectively. The conditional probability table is given as follows:

$$\begin{array}{l} P(X^1|Y^1), P(X^2|Y^1), \cdots, P(X^L|Y^1) \\ \vdots \\ P(X^1|Y^M), P(X^2|Y^M), \cdots, P(X^L|Y^M) \end{array}$$

where M and L are total numbers of the states of $Pa(x_i)$ and x_i, respectively.

3.2. Determination of graph structure

In this study, networks are determined by using K2 algorithm (Ben-Gal, 2007; Pearl & Russell, 2002) with K2Metric (Ben-Gal,

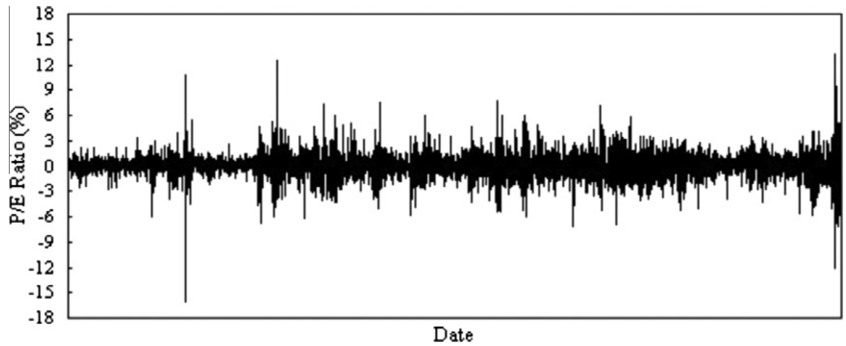

Fig. 1. P/E ratio of NIKKEI stock average.

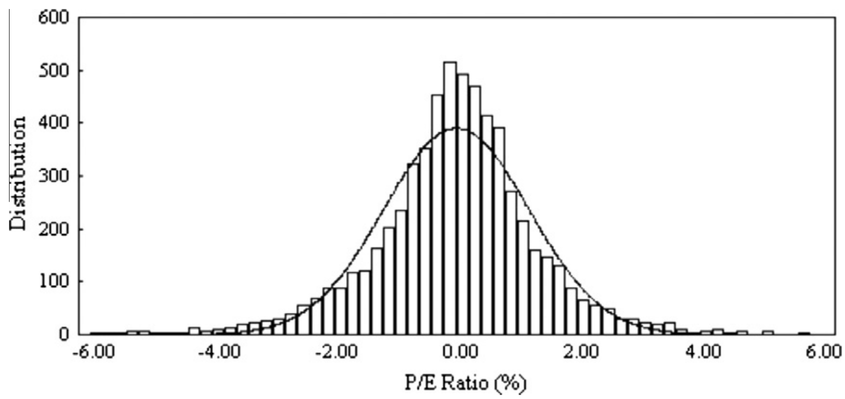

Fig. 2. P/E ratio frequency distribution of NIKKEI stock average.

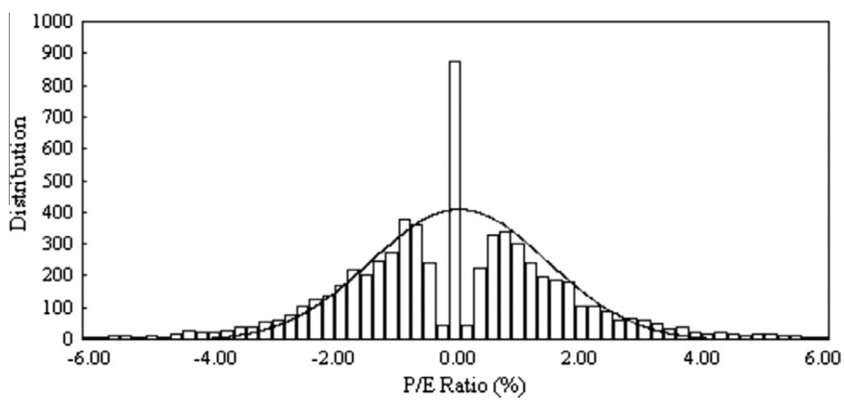

Fig. 3. P/E ratio frequency distribution of Toyota motor corporation stock price.

2007; Heckerman, Geiger, & Chickering, 1995; Pearl & Russell, 2002) as the estimator of the network.

K2Metric is given as follows (Cooper, 1992; Heckerman et al., 1995):

$$K2 = \prod_{i=1}^{N} \prod_{j=1}^{M} \frac{(L-1)!}{(N_{ij}+L-1)!} \prod_{k=1}^{L} N_{ijk}! \quad (8)$$

where

$$N_{ij} = \sum_{k=1}^{L} N_{ijk} \quad (9)$$

The notation N, L, and M denote total number of nodes, total numbers of states for x_i and $Pa(x_i)$, respectively. Besides, the notation N_{ijk} denotes the number of samples of $x_i = X^k$ when $Pa(x_i) = Y^j$.

K2 algorithm determines the network from the totally ordered set of the random variables.

The K2 algorithm is illustrated in Fig. 5 and summarized as follows:

1. $i = 1$.
2. Set the parents set $Pa(x_i)$ for the note x_i to an empty set.
3. Estimate K2 metric S_{best} of the network composed of x_i and $Pa(x_i)$.
4. For $j = i + 1, \ldots, N$,
   (a) Add x_j to $Pa(x_i)$.
   (b) Estimate K2 metric S of the network composed of x_i and $Pa(x_i)$.
   (c) Delete x_j from $Pa(x_i)$ if $S < S_{best}$.
5. $i = i + 1$.

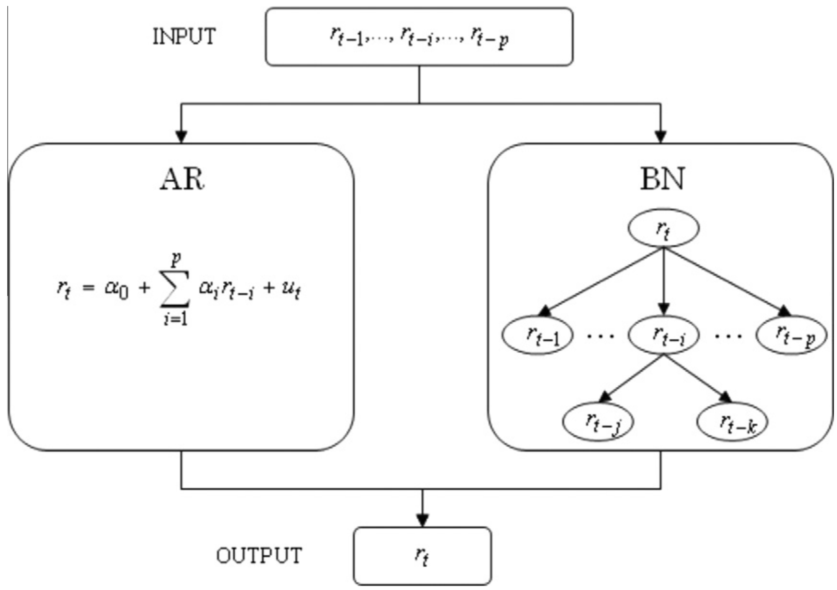

Fig. 4. Comparison of time-series and present algorithms.

6. Go to step 2 if $i \leqslant N$.

3.3. Probabilistic reasoning

When the evidence e of the random variable is given, the probability $P(x_i|e)$ is estimated by the marginalization with the conditional probability table (Pearl, 1988).

The marginalization algorithm gives the probability $P(x_i = X^l|e)$ as follows:

$$P(x_i = X^l|e) = \frac{\sum_{j=1, j\neq i}^{N} \sum_{x_j=X^1}^{X^L} P(x_1, \ldots, x_i = X^l, \ldots, x_N, e)}{\sum_{j=1}^{N} \sum_{x_j=X^1}^{X^L} P(x_1, \cdots, x_N, e)} \quad (10)$$

where the notation $\sum_{x_j=X^1}^{X^L}$ denotes the summation over all states $X^1, X^2, \ldots, X^L$ of the random variable x_j.

4. Present algorithm

4.1. Digitization of price–earnings ratio (P/E ratio)

The price–earnings ratio (P/E ratio) r_t is defined as follows.

$$r_t = (\ln P_t - \ln P_{t-1}) \times 100 \quad (11)$$

where the notation P_t denotes the closing stock price at time t.

The P/E ratio is digitized by the uniform clustering or Ward method. The set of digitized P/E ratio values is given as

$$\{r^1, r^2, \cdots, r^L\} \quad (12)$$

where the notation r^l ($l = 1, 2, \cdots, L$) and L denote the digitized P/E ratio values and its total number, respectively.

The notation C_l and c_l ($l = 1, 2, \cdots, L$) denote the cluster and its center. The cluster center is taken as the digitized values:

$$\{r^1, r^2, \cdots, r^L\} = \{c_1, c_2, \cdots, c_L\} \quad (13)$$

4.1.1. Uniform clustering

The uniform clustering algorithm divides clusters including same number of samples.

4.1.2. Ward method

In Ward method, the Euclid distances from samples to the cluster centers are minimized. The notation z, C_i and c_i denote the sample, the cluster and its center, respectively. The estimator is given as

$$D(C_i, C_j) = E(C_i \cup C_j) - E(C_i) - E(C_j) \quad (14)$$

$$E(C_i) = \sum_{z \in C_i} d(z, c_i)^2 \quad (15)$$

where the notation $d(z, c_i)$ denotes the Euclid distance between z and c_i.

4.2. Forecast of P/E ratio

For determining the network B by K2 algorithm, the total order of the random variable sets is necessary. The P/E ratio data are totally ordered according to the order of the time-series (Fig. 6).

Once the network B is determined, the P/E ratio r_t is forecasted so as to maximize the probability $P(r^l|B)$:

$$r_t = \arg\max_{r^l}(P(r^l|B)) \quad (16)$$

The P/E ratio forecast algorithm is summarized again as follows.

1. Digitize P/E ratio according to the algorithm in Section 4.1.
2. Determine Bayesian network B according to the algorithm in Section 3.
3. Predict P/E ratio by Eq. (16).

5. Numerical example

5.1. NKKEI stock average

The forecast of NIKKEI stock average (NIKKEI225) is considered as a first example. The NIKKEI225 P/E ratio and its distribution are shown in Figs. 1 and 2, respectively.

Bayesian network B is determined from the NIKKEI225 P/E ratio values from February 22nd 1985 to December 30th 2008. The network is applied for predicting the stock price from December 1st to December 30th, 2008. The accuracy is compared with traditional time-series forecast algorithms on the correlation coefficient (CC)

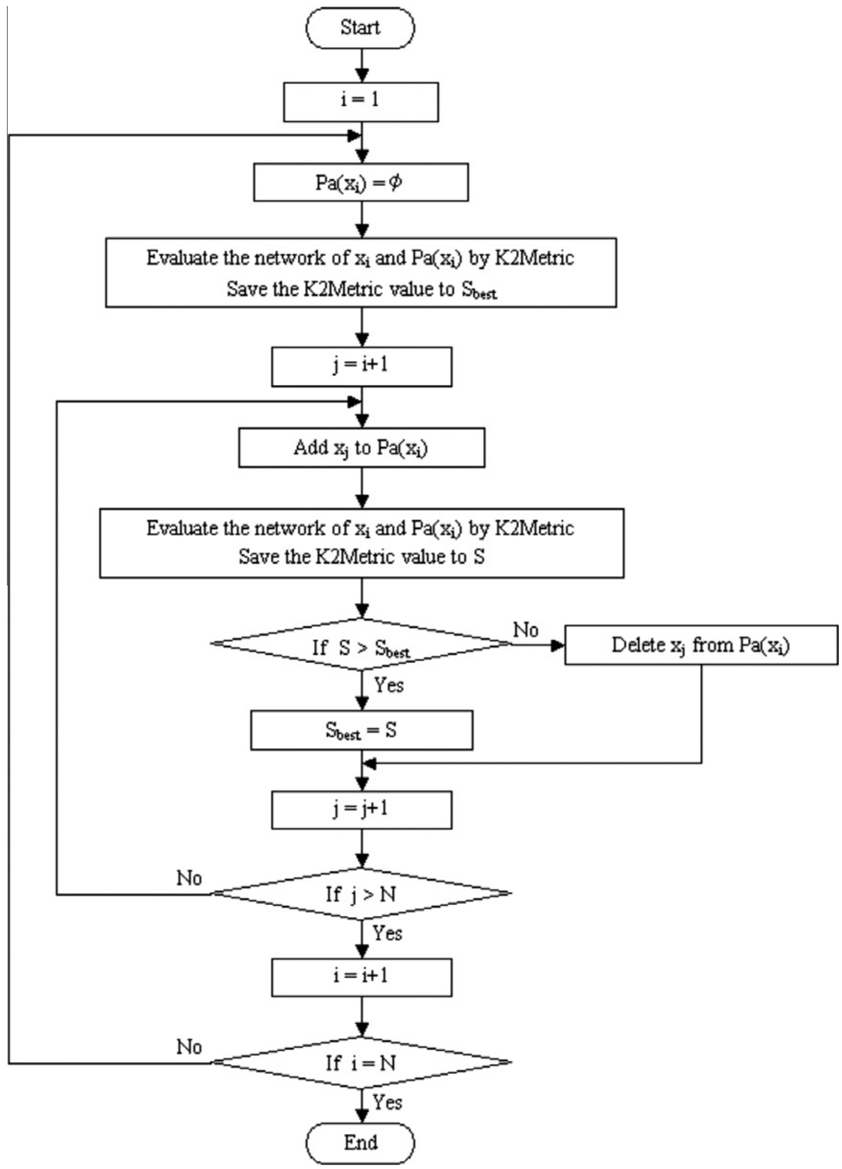

Fig. 5. K2 algorithm.

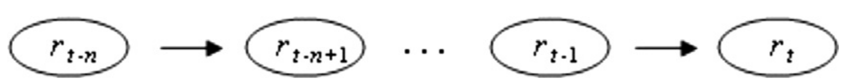

Fig. 6. Total order of P/E ratio.

Table 1
Classified section of NIKKEI stock average P/E ratio (uniform clustering).

Cluster	$(C_l)_{\min}$, $(C_l)_{\max}$	Number of samples	$c_l(r^j)$ (%)
C_1	[−16.14%, −1.12%)	980	−2.18
C_2	[−1.12%, −0.42%)	980	−0.73
C_3	[−0.42%, −0.00%]	980	−0.20
C_4	(+0.00%, 0.44%]	1020	0.22
C_5	(0.44%, 1.07%]	1020	0.72
C_6	(1.07%, 13.23%]	1020	2.04

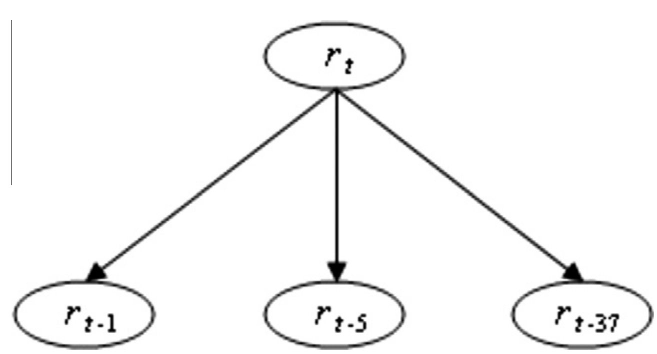

Fig. 7. Bayesian network for NIKKEI stock average (uniform clustering).

and the error (RMSE). The estimator CC and RMSE are estimated as follows:

Table 2
Classified section of NIKKEI stock average P/E ratio (Ward method).

Cluster	$(C_l)_{\min}$, $(C_l)_{\max}$	Number of samples	$c_l(r^l)$ (%)
C_1	[−16.14%, −3.03%)	150	−4.32
C_2	[−3.03%, −0.85%)	1123	−1.59
C_3	[−0.85%, −0.00%]	1667	−0.37
C_4	(+0.00%, 1.23%]	2207	0.52
C_5	(1.23%, 3.80%]	796	1.98
C_6	(3.80%, 13.23%]	57	5.47

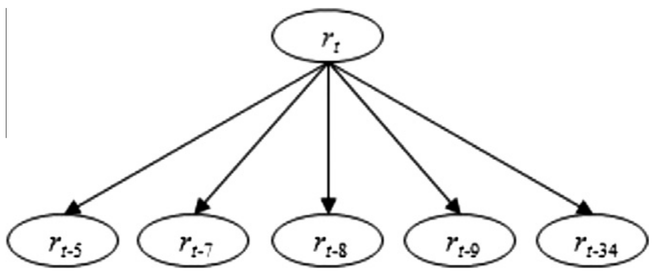

Fig. 8. Bayesian network for NIKKEI stock average (Ward method).

Table 3
Cluster center on different number of clusters for NIKKEI stock average (Ward method).

L	2	4	6	8	10
c_1	−1.04%	−4.32%	−4.32%	−11.66%	−11.66%
c_2	0.99%	−0.86%	−1.59%	−4.07%	−4.07%
c_3	–	0.52%	−0.37%	−1.59%	−2.10%
c_4	–	2.21%	0.52%	−0.38%	−1.16%
c_5	–	–	1.98%	0.52%	−0.38%
c_6	–	–	5.47%	1.57%	0.52%
c_7	–	–	–	2.63%	1.57%
c_8	–	–	–	5.47%	2.63%
c_9	–	–	–	–	5.01%
c_{10}	–	–	–	–	11.48%

$$\mathrm{CC} = \frac{\sum_{t=1}^{n}(r_t - \bar{r})(r'_t - \bar{r'})}{\sqrt{\sum_{t=1}^{n}(r_t - \bar{r})^2} \cdot \sqrt{\sum_{t=1}^{n}(r'_t - \bar{r'})^2}} \tag{17}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(r_t - r'_t)^2} \tag{18}$$

$$\bar{r} = \frac{1}{n}\sum_{t=1}^{n} r_t$$

$$\bar{r'} = \frac{1}{n}\sum_{t=1}^{n} r'_t$$

where the notation r_t and r'_t denote the actual and predicted stock prices, respectively. The notation n is the total number of data.

5.1.1. Comparison of clustering algorithms

Total number of clusters is specified as $L = 6$. The samples are clustered by an uniform clustering algorithm, which is listed in Table 1. The notation $(C_l)_{\min}$ and $(C_l)_{\max}$ denote the minimum and maximum values of the samples in the cluster C_l. The number of samples is the number of samples in the cluster C_l. The Bayesian network is determined from the data, which is shown in Fig. 7. Figure shows the P/E ratio r_t has dependency with one-day prior P/E ratio r_{t-1}, 5-days prior P/E ratio r_{t-5} and 37-days prior P/E ratio r_{t-37}.

Next, the samples are clustered by Ward method, which is listed in Table 2. The Bayesian network is determined from the data, which is shown in Fig. 8. This network is very different from that by uniform clustering because the P/E ratio r_t has the dependency on the 5-days prior P/E ratio r_{t-5}, the 7-days prior P/E ratio r_{t-7}, the 8-days prior P/E ratio r_{t-8}, the 9-days prior P/E ratio r_{t-9} and the 34-days prior P/E ratio r_{t-34}.

5.1.2. Comparison of cluster number

Next, the effect of the cluster number to the forecast accuracy is discussed. Samples are clustered by using Ward method into 2, 4, 6, 8 and 10 clusters. The forecast accuracy and the correlation coefficient with respect to the actual data are compared.

The cluster centers are listed in Table 3. The correlation coefficient and the errors are shown in Fig. 9. The label MaxErr, MinErr, RMSE, and CC denote the maximum error, the minimum error, the average error and the correlation coefficient, respectively. Figure shows that the error is smallest and the correlation coefficient is largest at the cluster number $L = 6$.

5.1.3. Comparison of forecast accuracy

The forecast accuracy is compared in Table 4. The label BN and BN2 denote the results which are predicted by Bayesian networks with a uniform clustering and the Ward method, respectively. The labels AR(2), MA(2), ARMA(2,2), ARCH(2,9) denote the results by AR model with $p = 2$, MA model with $q = 2$, ARMA model with $p = 2$ and $q = 2$ and ARCH model with $p = 2$ and $q = 9$, respectively. The parameters p and q are selected according to the algorithm in Section 2.1.5. In each case, 1000 simulations are performed and the average values are shown in the table.

Comparison of BN with traditional algorithms shows that the correlation coefficient of BN is higher than the traditional ones although their errors are almost similar. Besides, in the BN2, both error and correlation coefficient are improved against the other

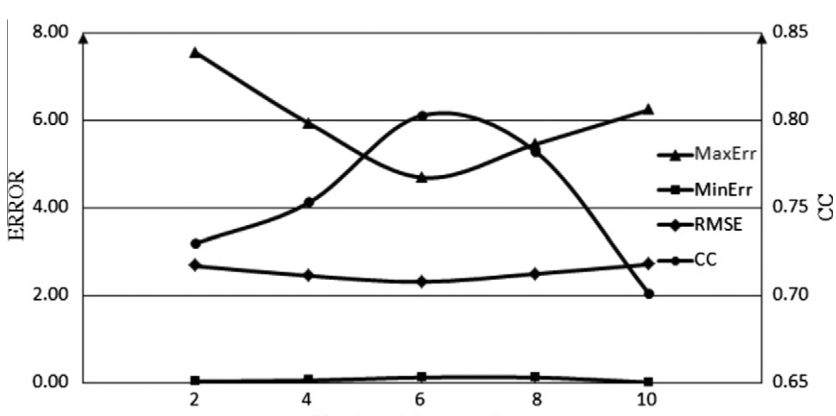

Fig. 9. Effect of number of digitized values to accuracy on NIKKEI stock average (Ward method).

Table 4
Comparison of predicted and actual stock prices (NIKKEI stock average).

	Max. error	Min. error	CC	RMSE
BN	5.5203	0.4172	0.7785	2.7521
BN2	4.6893	0.1162	0.8028	2.3065
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

Table 5
Classified section of Toyota motor corporation stock price P/E ratio (uniform clustering).

Cluster	$(C_l)_{\min}$, $(C_l)_{\max}$	Number of samples	$c_l(r^l)$ (%)
C_1	$[-21.13\%, -1.603\%)$	897	−2.84
C_2	$[-1.603\%, -0.777\%)$	897	−1.16
C_3	$[-0.777\%, -0.00\%)$	897	−0.50
C_4	$[+0.00\%, -0.00\%]$	720	0.00
C_5	$(+0.00\%, 0.80\%]$	863	0.50
C_6	$(0.80\%, 1.63\%]$	863	1.17
C_7	$(1.63\%, 16.25\%]$	863	3.08

algorithms including traditional ones and BN. It is shown that the importance of the clustering algorithm.

The frequency distributions of the P/E ratio are shown in Fig. 10. The figure is plotted with the P/E ratio as the horizontal axis and the frequency distribution as the vertical axis, respectively. In case of BN, the P/E ratio distribution is relatively even and thus, it is very far from the actual data. In case of BN2, the P/E ratio distribution is more similar to the actual data than that in BN. Therefore, the BN2-forecast accuracy is better than the BN-forecast accuracy.

5.2. Toyota motor corporation

The forecast of the P/E ratio of the Toyota motor corporation stock price is considered as a second example. The P/E ratio frequency distribution of the Toyota motor corporation stock price is shown in Fig. 3. It is shown that the distribution is far from the normal distribution and especially, the distribution near P/E ratio = 0% is much larger than the normal distribution.

Bayesian network B is determined from the P/E ratio values of the Toyota motor corporation stock price from February 22nd 1985 to December 30th 2008. The network is applied for predicting the P/E ratio from December 1st to December 30th, 2008. The accuracy is compared with traditional time-series forecast algorithms on the correlation coefficient (CC) and the root mean square error (RMSE).

5.2.1. Comparison of clustering algorithms

The number clusters is specified as $L = 7$. First, the uniform clustering is applied to the samples. The digitized numbers are listed in Table 5. The notation $(C_l)_{\min}$ and $(C_l)_{\max}$ denote the minimum and maximum values of the samples in the cluster C_l. The number of samples is the number of samples in the cluster C_l. The Bayesian network is determined from the data, which is shown in Fig. 11. Figure shows the P/E ratio r_t has dependency with one-day prior P/E ratio r_{t-1} alone.

Next, the samples are clustered by Ward method, which is listed in Table 6. The Bayesian network is determined from the data, which is shown in Fig. 12. This network is very different from that by uniform clustering because the P/E ratio r_t has the dependency on the one-day prior P/E ratio r_{t-1}, the 5-days prior P/E ratio r_{t-5}, the 8-days prior P/E ratio r_{t-6}, the 7-days prior P/E ratio r_{t-7}, the 8-days prior P/E ratio r_{t-8} and the 9-days prior P/E ratio r_{t-18}.

5.2.2. Comparison of cluster number

Next, the effect of the cluster number to the forecast accuracy is discussed. Samples are clustered by using Ward method into 3, 5, 7, 9 and 11 clusters. The forecast accuracy and the correlation coefficient with respect to the actual data are compared.

The cluster centers are listed in Table 7. The correlation coefficient and the errors are shown in Fig. 13. The label MaxErr, MinErr, RMSE, and CC denote the maximum error, the minimum error, the average error and the correlation coefficient, respectively. Figure shows that the error is smallest and the correlation coefficient is largest at the cluster number $L = 7$.

5.2.3. Comparison of forecast accuracy

The forecast accuracy is compared in Table 8. The label BN and BN2 denote the results which are predicted by Bayesian networks with a uniform clustering and the Ward method, respectively. The labels AR(9), MA(6), ARMA(9,6) and ARCH(9,6) denote the results by AR model with $p = 9$, MA model with $q = 6$, ARMA model with $p = 9$ and $q = 6$ and ARCH model with $p = 9$ and $q = 6$, respectively. The parameters p and q are selected according to the algorithm in Section 2.1.5. In each case, 1000 simulations are performed and the average values are shown in the table.

Table 8 shows that, in BN, AR, MA, ARMA, ARCH, the correlation coefficient is smaller than 0.5 and the root mean square error is greater than 4. However, in BN2, the correlation coefficient and

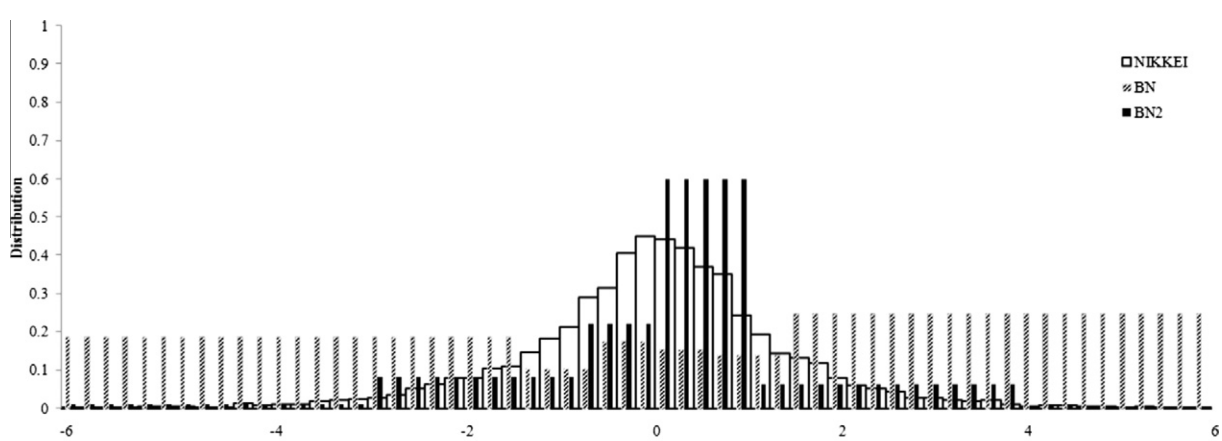

Fig. 10. Comparison of P/E ratio frequency distributions (NIKKEI stock average).

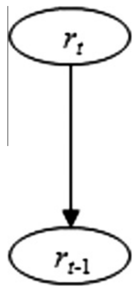

Fig. 11. Bayesian network for P/E ratio of Toyota motor corporation (uniform clustering).

Table 6
Classifying section of Toyota motor corporation stock price P/E ratio (Ward method).

Cluster	$(C_l)_{\min}$, $(C_l)_{\max}$	Number of samples	$c_l(r^l)$ (%)
C_1	[−21.13%, −3.69%)	156	−5.31
C_2	[−3.69%, −1.25%)	1062	−2.05
C_3	[−1.25%, −0.00%)	1473	−0.69
C_4	[+0.00%, −0.00%]	720	0.00
C_5	(+0.00%, 1.31%]	1456	0.72
C_6	(1.31%, 3.98%]	970	2.16
C_7	(3.98%, 16.25%]	163	5.85

Table 7
Discrete number on different number of clusters for Toyota motor corporation stock price (Ward method).

L	3	5	7	9	11
c_1	−1.51%	−5.31%	−5.31%	−8.20%	−8.20%
c_2	0.00%	−1.26%	−2.05%	−4.45%	−4.45%
c_3	1.59%	0.00%	−0.69%	−2.05%	−2.68%
c_4	–	1.29%	0.00%	−0.69%	−1.64%
c_5	–	5.85%	0.72%	0.00%	−0.69%
c_6	–	–	2.16%	0.72%	0.00%
c_7	–	–	5.85%	2.16%	0.72%
c_8	–	–	–	5.34%	1.64%
c_9	–	–	–	10.80%	2.74%
c_{10}	–	–	–	–	5.34%
c_{11}	–	–	–	–	10.80%

Table 8
Comparison of predicted and actual stock prices (Toyota motor corporation).

	Max. error	Min. error	CC	RMSE
BN	12.1722	0.1583	0.4849	5.0698
BN2	7.1762	0.1808	0.6284	3.2669
AR(9)	10.3189	0.2119	0.4689	4.0866
MA(6)	10.2584	0.0242	0.4679	4.1016
ARMA(9,6)	10.6026	0.0539	0.4457	4.1434
ARCH(9,9)	10.2630	0.1632	0.4692	4.0669

the root mean square error are 0.6284 and 3.2669, respectively. They are much better than the others.

The frequency distributions of the P/E ratio are shown in Fig. 14. The figure is plotted with the P/E ratio as the horizontal axis and the frequency distribution as the vertical axis, respectively. The labels TOYOTA, BN and BN2 denote the actual distribution, the distributions in BN and BN2, respectively. While, in case of BN, the distribution is relatively even, the distribution in BN2 is much more similar to the actual data except for that near 0.

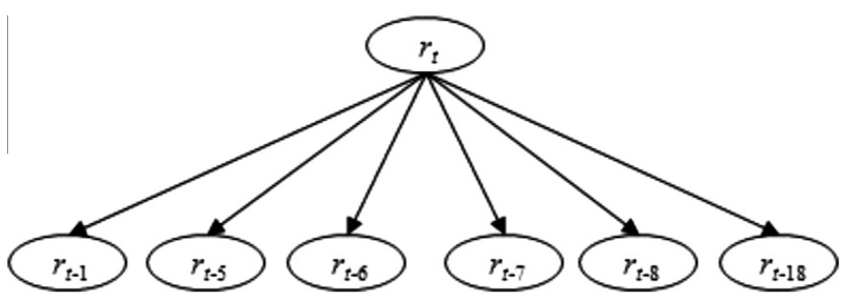

Fig. 12. Bayesian network for Toyota motor corporation stock price P/E ratio (Ward method).

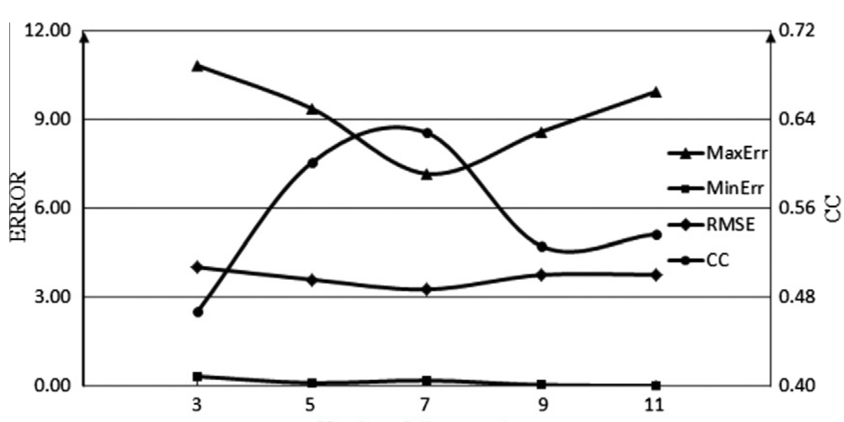

Fig. 13. Effect of number of discrete values to accuracy (Toyota motor corporation stock price P/E ratio).

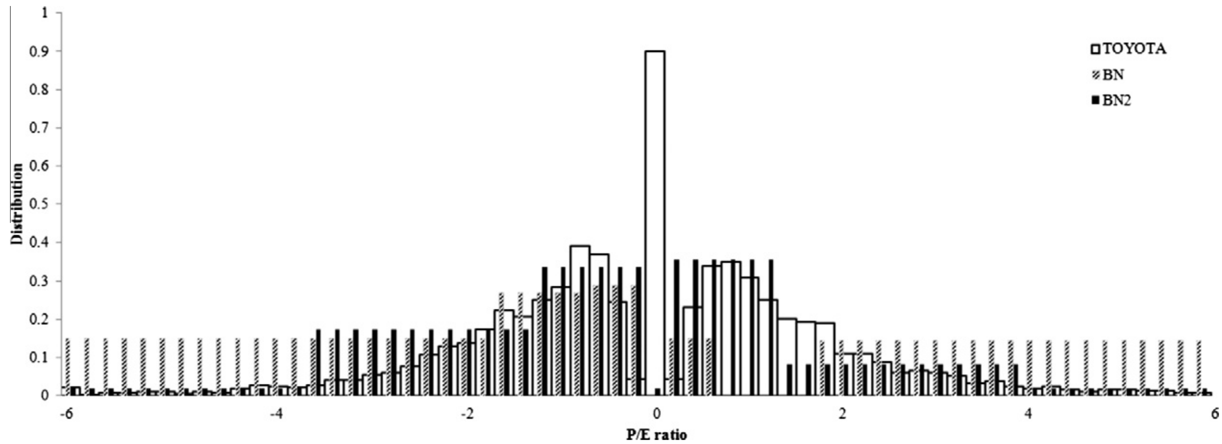

Fig. 14. Comparison of P/E ratio frequency distributions (Toyota motor corporation).

6. Conclusions

The P/E ratio forecast algorithm by using Bayesian network was presented in this study. The P/E ratio values are digitized by clustering the P/E ratio frequency distribution by the uniform clustering or Ward method. Bayesian network for dependency among previous P/E ratio distribution is determined from the digitized P/E ratio values. The forecast accuracy and the correlation coefficient with respect to the actual stock price are compared with the traditional time-series forecast algorithms such as AR, MA, ARMA and ARCH models.

Through the numerical example of Nikkei stock average and Toyota motor corporation stock price, the present algorithm using the uniform clustering shows the similar accuracy and the better correlation coefficient against to the time-series forecast algorithms. And in the present algorithm using Ward method, the computational accuracy is improved by 15% (NIKKEI stock average) and 20% (Toyota motor corporation stock price) against the traditional ones.

In the next study, we would like to improve the present method still more by developing the P/E ratio digitizing algorithm.

References

Ben-Gal, I. (2007). Bayesian networks. In F. Ruggeri, R. S. Kennett, & F. W. Faltin (Eds.). *Encyclopedia of statistics in quality and reliability* (Vol. 1). John Wiley & Sons.

Bollerslevb, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*, 307–327.

Box, G., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control*. Prentice-Hall.

Brockwell, P., & Davis, R. (2009). *Time series: Theory and methods*. Springer.

Cooper, G. F. E. H. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*(4), 309–347.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance, 48*(5), 1749–1778.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*, 197–243.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica, 59*, 347–370.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. *Representation and reasoning series*. Morgan Kaufmann.

Pearl, J., & Russell, S. (2002). Bayesian networks. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 157–160). MIT Press.

Takayasu, H. (2006). *Practical fruits of econophysics*. Springer.