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:

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

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:

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

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:

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

2.1.4. ARCH model

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

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

The error term u_t is given as follows:

$$u_t = \sigma_t z_t \tag{5}$$

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

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

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:

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

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:

$$P(X^1|Y^1), P(X^2|Y^1), \cdots, P(X^L|Y^1)$$
$$\vdots$$
$$P(X^1|Y^M), P(X^2|Y^M), \cdots, P(X^L|Y^M)$$

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,