Neurocomputing 83 (2012) 136–145

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

A novel text mining approach to financial time series forecasting

Baohua Wang [a,b,*], Hejiao Huang [a], Xiaolong Wang [a]

[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
[b] College of Mathematics and Computational Science, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 15 January 2011
Received in revised form
9 September 2011
Accepted 4 December 2011
Communicated by P. Zhang
Available online 30 December 2011

Keywords:
Financial time series forecasting
ARIMA
Support vector regression
Market sentiment

ABSTRACT

Financial time series forecasting has become a challenge because it is noisy, non-stationary and chaotic. Most of the existing forecasting models for this problem do not take market sentiment into consideration. To overcome this limitation, motivated by the fact that market sentiment contains some useful forecasting information, this paper uses textual information to aid the financial time series forecasting and presents a novel text mining approach via combining ARIMA and SVR (Support Vector Regression) to forecasting. The approach contains three steps: representing textual data as feature vectors, using ARIMA to analyze the linear part and developing a SVR model based only on textual feature vector to model the nonlinear part. To verify the effectiveness of the proposed approach, quarterly ROEs (Return of Equity) of six security companies are chosen as the forecasting targets. Comparing with some existing state-of-the-art models, the proposed approach gives superior results. It indicates that the proposed model that uses additional market sentiment provides a promising alternative to financial time series prediction.

1. Introduction

A time series is a sequence of numerical data points recorded sequentially in time. Time series forecasting aims to develop a model describing its underlying relationship by collecting and analyzing the past observations; the future values of the variable are then predicted based on the established model. In particular, financial time series forecasting, which is an important issue in investment and financial decision-making, has become an active research area and has drawn considerable attentions. However, financial time series forecasting is regarded as one of the most challenging applications of modern time series forecasting because it is inherently noisy, non-stationary and deterministically chaotic [1].

Numerous attempts have been made to improve the accuracy of prediction. Introduced by Box and Jenkins, the ARIMA model [2] has been one of the most popular approaches to time series forecasting. The ARIMA model, which tries to find the optimal number of previous values of the dependent variable and random shocks, has the fundamental impact on the time series analysis and forecasting applications. However, its major limitation is the pre-assumed linear form of the model. In addition, due to data limitation, it requires at least fifty, or preferably one hundred and more sets of historical data to yield desired results [3]. Methods are needed to deal with the nonlinear pattern and the situations where only small quantities of historical data are available. Our model, therefore, is developed to achieve objectives in situations with small quantities of historical data available.

To complement each other and make use of each model's unique feature, some hybrid models [3–5] have been proposed. The results of hybrid models have suggested that combination of forecasts from more than one model often leads to improved forecasting performance [6–8]. The main reason is that it is possible for a data generation process to switch its structure over the observation period between a linear and nonlinear structure [9,10]. In neural network forecasting research, a number of combining schemes have been proposed. Terui et al. [10] have presented a linear and nonlinear time series model for forecasting the US monthly employment rate and production indices. Luxhoj et al. [11] have presented a hybrid econometric-neural network modeling approach for sales forecasting. Zhang [12] has proposed a hybrid ARIMA and neural network model to improve time series forecasting accuracy. Khashei et al. [3] have combined the ARIMA models with Artificial Neural Networks (ANNs) and Fuzzy logic in order to overcome the linear and data limitations of ARIMA models. Their results have shown the superiority of the combined forecasting model.

Recently, a novel neural network algorithm, called support vector machines (SVM), was developed by Vapnik et al. [13]. SVM implements the structural risk minimization principle which seeks to minimize an upper bound of the generalization error rather than the training error. With the introduction of Vapnik's ε-insensitive loss function, the regression model of SVMs, called

* Corresponding author at: School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China. Tel.: +86 755 26538408; fax: +86 755 26534791.
E-mail address: bhwang@szu.edu.cn (B. Wang).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.12.013

support vector regression (SVR), has been extended to solve nonlinear regression estimation problems [14]. It has been applied to financial time series problems and has shown a quite good performance [15–17]. However, real-world financial time series often contain both linear and nonlinear patterns, ARIMA and SVM cannot achieve the best performance in every situation [18]. So SVM has been introduced to the hybrid model for time series forecasting since, theoretically, the SVM algorithm is superior to the ANN model for this problem [19–21]. Pai et al. [4] have presented a hybrid ARIMA and SVM model for stock price forecasting. Chen et al. [5] have proposed a hybrid methodology that exploits the unique strength of the seasonal autoregressive integrated moving average (SARIMA) model and the SVM model in forecasting seasonal time series. Ni et al. [22] describes a hybrid model formed by a mixture of various regressive neural network models, including support vector regressions, and a selected set of influential trading indicators are utilized, for modeling and prediction of foreign exchange rate time series. Their results show that the SVM-combined forecasts outperform the individual forecasts.

Nevertheless, existing hybrid models are solely quantitative data based models, without considering the other factors, such as market sentiment. This may reduce the forecasting accuracy. Textual data may be available to users in real-world financial time series forecasting applications and it may contain more useful information than the financial ratios [23]. For example, industrial analysts usually uncover the indications and hints about future financial performance by reading not only financial numbers but also the textual part of a company's financial reports and make "professional guesses." As shown in [23], the quantitative part of a report only reflects the past performance of a company, but at the same time, the qualitative part of a report holds some message about the future company performance. Recently, some efforts have been made to build forecasting models upon textual data. Gidófalvi et al. [24] gathered over 5000 financial news articles about 12 stocks and demonstrated that there exists a weak ability to predict the direction of a security before the market corrects itself to equilibrium. Wuthrich et al. [25] predict stock markets using information contained in articles published on the Web, improved predictions are presented by exploiting textual information in addition to numeric time series data. Fung et al. [26] have proposed a systematic framework for mining multiple time series concurrently by using textual documents as the source of prediction and indicate that their methodology is both theoretically and practically possible. Kloptchenko et al. [23] have proven that some future changes in financial performance can be anticipated by analyzing the texts from quarterly reports. Before a dramatic change occurs in company financial performance, a change can be seen in the written style of a financial report and the tone tends to be closer to the next quarterly performance. Schumaker [27] has developed a model which uses article terms and the stock price at the time when the article releases and shows that it is better able to capture stock price movements and further bolsters the idea of weak short-term predictability.

However, as far as we know, there has not been any hybrid model that uses additional market sentiment for financial time series forecasting yet. The main reason for this is that it is difficult to extract relevant information from these unstructured text data. Text mining methods are needed to study the hidden indications about the future financial performance from the textual data. Feature selection and regression-function are two main tasks in the text mining literature. The task of feature selection is to present the textual data as appropriate weighted features. Regression-function is the function trained to extract quantitative information from these features. In this paper, each text is represented as a feature vector in a high-dimensional term frequency/inverse document frequency (TF-IDF) vector space. Since there are many successful results of utilizing SVR in nonlinear prediction, SVR is used as the regression-function in this research work.

In this way, the proposed method consists of three steps. In the first step, text mining techniques are exploited to represent each text as a feature vector in a high-dimensional TF-IDF vector space. Since real-world financial time series often contain both linear and nonlinear patterns, an ARIMA model analyzes the linear part of the problem in the second step; the residuals which contain nonlinear information can be then calculated. In the third step, a SVR model built only upon the textual feature vector is developed to model the residuals from the ARIMA model, that is, the SVR model can give the prediction of the future observation's residual after training.

The main contributions of this paper are two-fold: (1) market sentiment and text mining techniques are introduced to solve the financial time series problem, thus precision can be improved by combining heterogeneous data; and (2) a hybrid model which combines the traditional ARIMA model with SVR method is proposed to extract useful information from the market sentiment, it provides a new promising alternative to financial time series prediction.

The rest of this paper is organized as follows. The next section describes the proposed methodology, and then experiment and evaluation results are explained and discussed in Section 3. Finally, Section 4 draws the conclusions.

2. Proposed financial time series forecasting method

2.1. Definitions

Assume the time series is $Y=\{y_1,\ldots y_t,\ldots,y_N\}$ where y_t is the value at time t, and there exists the market sentiment set $T=\{T_1,\ldots,T_t,\ldots,T_N\}$ where T_t is the textual data available at time t, N is the number of sample observations.

2.2. Proposed approach

Motivated by the fact that market sentiment contains some useful forecasting information, this paper develops a market sentiment based model via combining ARIMA and SVR for financial time series forecasting. Details of the approach are discussed as follows.

Forecasting information may not be explicitly shown in textual data but rather encapsulated in forward-looking statements, so text mining technique is used in this study. Feature selection is one of the main issues in the text mining literature. There are many methods available to feature selection. One of the more common methods is to apply a vector representation where terms are indexed and then weighted. The type of representation which is the most popular in the text representation literature is known as the "bag-of-words" approach [28]. So each T_t is represented using bag-of-words technique in this study. Bag-of-words text representation approach identifies the important terms and each feature corresponds to an important term, which is found in the training corpus. To reduce the dimension of the original feature space, the terms which occurred three or more times are selected to form a feature vector [29]. The terms extracted from the text are represented in weights which are specified with TF-IDF measure [30]. TF-IDF weight $x_{i,t}$ for term k_i in T_t is defined by

$$x_{i,t}=TF_{i,t}\times IDF_i, \tag{1}$$

where $TF_{i,t}$ represents the normalized frequency of term k_i in text T_t, IDF_i represents the inverse document frequency. Thus, the textual data T_t is represented as a vector $X_t=(x_{1,t},\ldots,x_{i,t},\ldots,x_{V,t})$ where V is the size of the vocabulary.

Considering that a time series $Y=\{y_1,\ldots y_t,\ldots,y_N\}$ is composed of the linear part and the nonlinear part, y_t can be represented as follows [12]:

$$y_t=l_t+nl_t, \tag{2}$$

where l_t is the linear part and nl_t is the nonlinear part. l_t and nl_t are estimated from the numerical time series and textual data respectively.

ARIMA model is first utilized to learn the linear component l_t. In an ARIMA model, it is a mix between autoregressive and moving average models for the description of time series, expressed as follows:

$$l_t=\theta_0+\varphi_1 l_{t-1}+\cdots+\varphi_p l_{t-p}+e_t-\theta_1 e_{t-1}-\cdots-\theta_q e_{t-q}, \tag{3}$$

where l_t is the actual value and e_t is the random error at time t, φ_i and θ_j are the coefficients, p and q are integers which are greater than or equal to zero and refer to the order of the autoregressive and moving average parts of the model respectively. Basically, this method follows an iterative three phases: model identification, parameter estimation and diagnostic checking.

Let nl_t denote the residual at time t obtained from the ARIMA model, then

$$nl_t=y_t-\hat{l}_t, \tag{4}$$

where $\hat{l}_t$ is the forecast value of the ARIMA model at time t. The residuals $nl_2,\ldots,nl_t,\ldots,nl_N$ from the linear model will contain only the nonlinear relationship.

Given the feature vectors and the residuals, the problem of estimating the residuals can be obviously formulated as a supervised learning problem in the text mining literature. As mentioned above, the feature vectors are TF-IDF weighted vectors, which represent the term-frequency/inverse document frequency. Forecasting information may be encapsulated in some important terms' frequency. For example, it is obviously true that positive words may have higher frequency than negative words when financial performance turns better. So a non-linear regression-function is needed to learn the relation between the term-frequency and residuals. SVM is the most accurate classifier and the fastest to train in text categorization. Since the SVR is an adaptation of SVM and it has been successfully applied in different problems of the times series prediction, SVR is utilized in this paper for the regression-function.

Table 1
Time period and size of the six companies.

Company	Time period (size)
Merchants Bank	1/1/2002–12/31/2009 (32)
VANKE	1/1/2002-12/31/2009 (32)
SUNING APPLIANCE	1/1/2004–3/31/2010 (25)
GE	1/1/1994-12/31/2010 (68)
Johnson&Johnson	1/1/1994–12/31/2010 (68)
McDonald's	1/1/1994–12/31/2010 (68)

So the residuals are then modeled by the SVR built only upon textual feature vector $X_1,\ldots,X_t,\ldots,X_{N-1}$. The hypothesis here is that the market sentiment X_{t-1} holds some message about future performance and has ability to predict this observation's residual value nl_t which obviously is non-linear. The SVM regression function is formulated as follows [14]:

$$nl_t=f(X_{t-1})=w\phi(X_{t-1})+b, \tag{5}$$

where ϕ is a nonlinear mapping, which maps the original input space to a high dimensional feature space. The basic idea of SVR is to map the input space into a higher dimensional feature space via a nonlinear mapping and then form a linear function such that it deviates least from the training data according to the loss function while at the same time it is as flat as possible. For the estimation of the coefficients w and b, one can refer to [13]. Gaussian kernel function $K(x,y)=\exp(-(x-y)^2/(2\delta)^2)$ is used in this paper.

In summary, the error term nl_t from ARIMA and the textual feature vector X_{t-1}, which are used as target value, feature vector respectively, are fed into Eq.(5) where textual data's predictability is trained and then the SVR model can give the prediction of residuals after training. Therefore, the combined forecast is

$$\hat{y}_t=\hat{l}_t+\hat{nl}_t. \tag{6}$$

Notably, $\hat{nl}_t$ is the forecast value of Eq. (4). The model exploits the unique feature and strength of market sentiment in determining nonlinear patterns. Thus, it can be advantageous to improve the overall modeling and forecasting performance.

2.3. Algorithm design

From the above analysis, the algorithm is designed in details as follows:

Step 1: Represent the textual data $T=\{T_1,\ldots,T_t,\ldots,T_N\}$ with bag-of-words text representation approach mentioned in Section 2.2. Let vector $X_t=(x_{1,t},\ldots,x_{i,t},\ldots,x_{V,t})$ represent the textual data T_t.

Step 2: Develop the ARIMA model based on the given financial time series $Y=\{y_1,\ldots y_t,\ldots,y_N\}$.

Step 3: Calculate residual value nl_t defined in Eq. (4) from the ARIMA model.

Table 3
Performance comparison on Merchants Bank when (a) $N=20$, (b) $N=24$, (c) $N=28$.

Metrics	Model 1	Model 2	Our model
(a)			
MAE	1.68	1.78	1.66
MAPE	27.81	29.83	27.67
RMSE	2.02	2.09	2.00
(b)			
MAE	1.28	1.21	0.94
MAPE	21.14	22.63	15.93
RMSE	1.57	1.42	1.20
(c)			
MAE	0.79	0.94	0.39
MAPE	15.71	18.88	7.77
RMSE	0.86	1.07	0.43

Table 2
Parameters of different models on Merchants Bank.

N	Model 1	Model 2	Our model
20	ARIMA (4,0,0)	$c=2.0$, $\gamma=512.0$, $\varepsilon=0.0625$	$c=0.25$, $\gamma=0.007813$, $\varepsilon=0.5$
24	ARIMA (1,0,0)	$c=1024$, $\gamma=4.0$, $\varepsilon=1.0$	$c=4.0$, $\gamma=0.000977$, $\varepsilon=0.5$
28	ARIMA (1,0,0)	$c=4.0$, $\gamma=8.0$, $\varepsilon=0.5$	$c=4.0$, $\gamma=0.000977$, $\varepsilon=0.5$

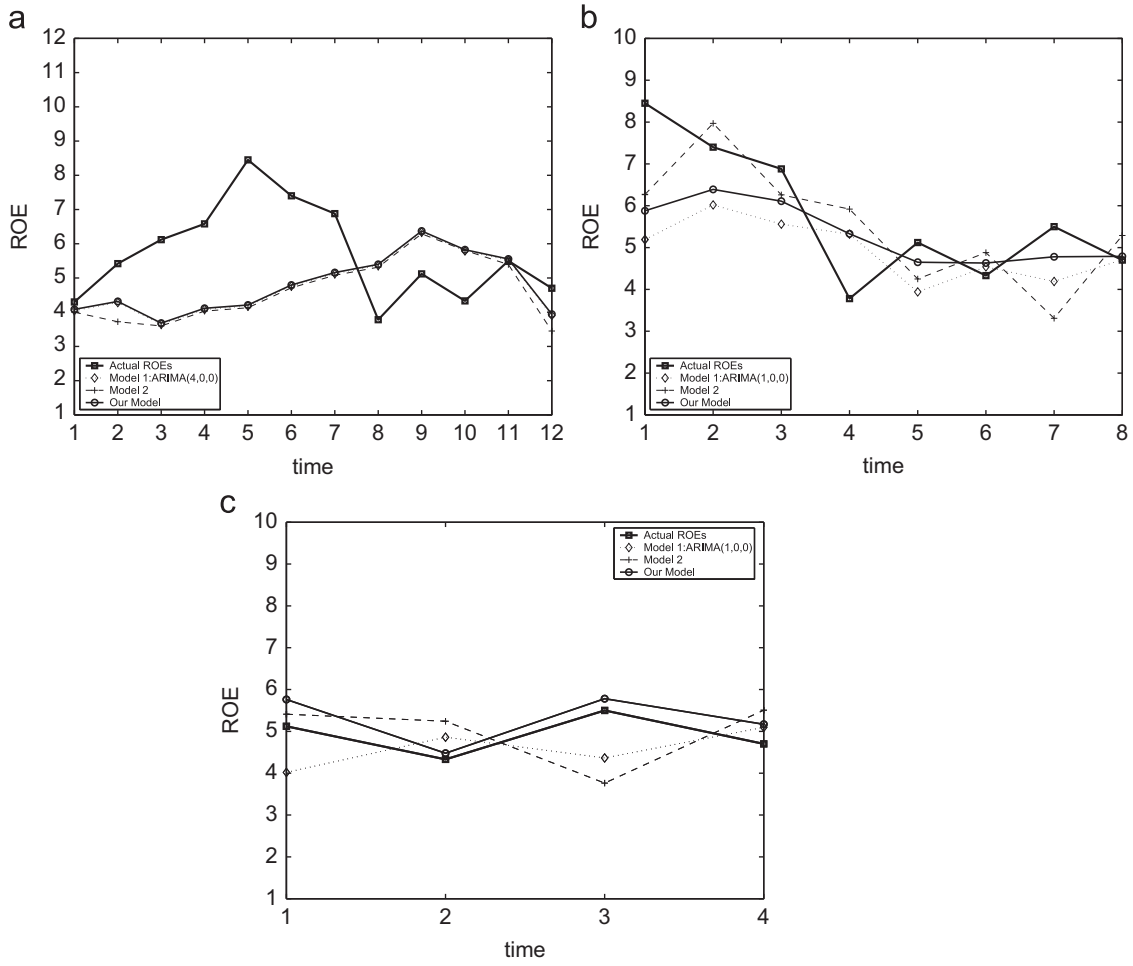

Fig. 1. Actual and predicted ROEs of merchants bank when (a) $N=20$, (b) $N=24$, (c) $N=28$.

Table 4
Parameters of different models on VANKE.

N	Model 1	Model 2	Our model
20	ARIMA (4,1,0)	$c=8.0$, $\gamma=8.0$, $\varepsilon=1.0$	$c=0.25$, $\gamma=0.03125$, $\varepsilon=0.25$
24	ARIMA (4,0,0)	$c=16.0$, $\gamma=8.0$, $\varepsilon=0.00098$	$c=0.0039063$, $\gamma=1.0$, $\varepsilon=0.003906$
28	ARIMA (4,0,0)	$c=1024.0$, $\gamma=4.0$, $\varepsilon=1.0$	$c=2.0$, $\gamma=0.000977$, $\varepsilon=1.0$

Step 4: The residuals series $nl_2,\ldots,nl_t,\ldots,nl_N$ are then modeled by a SVR model built only upon textual feature vector $X_1,\ldots,X_t,\ldots,X_{N-1}$, that is, a SVR model, which uses nl_t, X_{t-1} as the target value and feature vector respectively, is developed.

Step 5: To predict the time series value y_{N+1}, calculate $\hat{l}_{N+1}$ based on the ARIMA model developed in Step 2, calculate $\hat{nl}_{N+1}$ based on the SVR model developed in Step 4 upon the feature vector X_N. Let $\hat{y}_{N+1}=\hat{l}_{N+1}+\hat{nl}_{N+1}$, hence, y_{N+1}'s forecasting value $\hat{y}_{N+1}$ is obtained.

3. Experimental results

This section reports the experimental results of the proposed algorithm applied to quarterly Return on Equity (ROE) time series of six security companies.

3.1. Data sets

Six security companies' quarterly and annual reports are collected and their ROEs are selected as financial time series. Three companies which won the Chinese Best Investor Relationship Award in 2009 are: Merchants Bank (CHINA MERCHANTS BANK), VANKE (CHINA VANKE CO., LTD) and SUNING APPLIANCE (SUNING APPLIANCE CO., LTD) and three companies which often won the Best Investor Relationship Award in US is: GE (GENERAL ELECTRIC COMPANY), Johnson&Johnson (JOHNSON&JOHNSON COMPANY) and McDonald's (McDONALD's CORPORATION). Quarterly ROEs are selected as financial time series because ROE reveals the rate at which shareholders are earning income on their shares and it is one of the most important profitability indicators for a security company. Their corresponding time periods of financial reports and total numbers of ROEs time series, which are available on their official websites at the time when this paper is written, are listed in Table 1. It can be found that only small quantities of historical data are available.

The *Management's Analysis and Discussion* in the quarterly or annual reports are used as textual analysis data, since these textual parts might reveal some things that the companies may not wish to announce directly to their outside audience. By performing these actions, this study gathers 32, 32, 25, 68, 68 and 68 texts from the six companies respectively. To perform our

Chinese textual analysis, ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) [31], which performs Chinese word segment as well as POS tagging, is selected. ICTCLAS is a HHMM-based Chinese lexical analyzer and was found to have an approximately 90% *F*-measure for both precision and recall, which is comparable to other tools. R text-mining package tm [32] is used for English words processing. This pool of texts was analyzed using bag-of-words representation and retained only those terms that appeared three or more times.

Table 5
Performance comparison on VANKE when (a) $N=20$, (b) $N=24$, (c) $N=28$.

Metrics	Model 1	Model 2	Our model
(a)			
MAE	1.13	1.34	1.08
MAPE	37.73	43.98	36.61
RMSE	1.74	1.78	1.65
(b)			
MAE	0.73	0.78	0.69
MAPE	44.30	52.41	46.84
RMSE	0.84	0.97	0.77
(c)			
MAE	0.75	1.28	0.72
MAPE	27.36	48.20	22.67
RMSE	0.78	1.31	0.79

The experiment was conducted comparing with the other models: pure ARIMA model (Model 1), hybrid ARIMA and SVM using only numbers (Model 2).

In Model 1, the ARIMA model is implemented via EVIEW system. The most appropriate ARIMA model for different companies was based on autocorrelation and partial autocorrelation.

In Model 2, a hybrid model is built only upon numbers as [4]; ARIMA served as a preprocessor to filter the linear pattern of data sets. Then, the error terms from ARIMA were fed into the SVM in the hybrid models. The SVM model using only numbers was conducted to reduce the error function from the ARIMA.

Three parameters – γ, ε and C in SVM models in both Model 2 and our model – were adjusted based on the training sets. The parameter sets with the lowest values of MSE were selected for use. For the SVM calculations, the LIBSVM software system [33] was used.

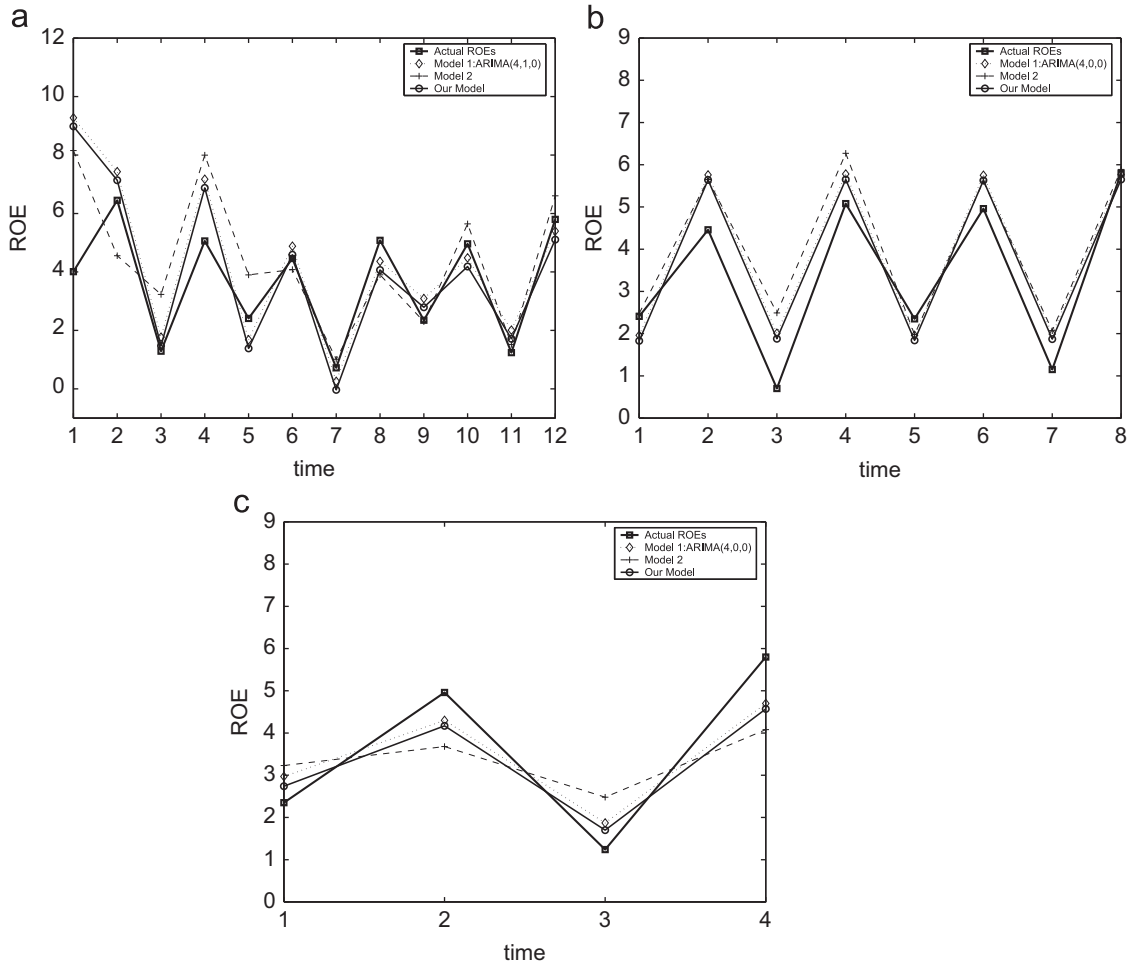

Fig. 2. Actual and predicted ROEs of VANKE when (a) $N=20$, (b) $N=24$, (c) $N=28$.

Table 6
Parameters of different models on SUNNING APPLIANCE.

N	Model 1	Model 2	Our model
20	ARIMA (0,1,1)	$c=32.0$, $\gamma=0.5$, $\varepsilon=2.0$	$c=8.0$, $\gamma=0.003906$, $\varepsilon=0.000977$

Table 7
Performance comparison on SUNNING APPLIANCE when $N=20$.

Metrics	Model 1	Model 2	Our Model
MAE	1.49	1.44	1.38
MAPE	21.70	23.23	20.53
RMSE	1.68	1.92	1.47

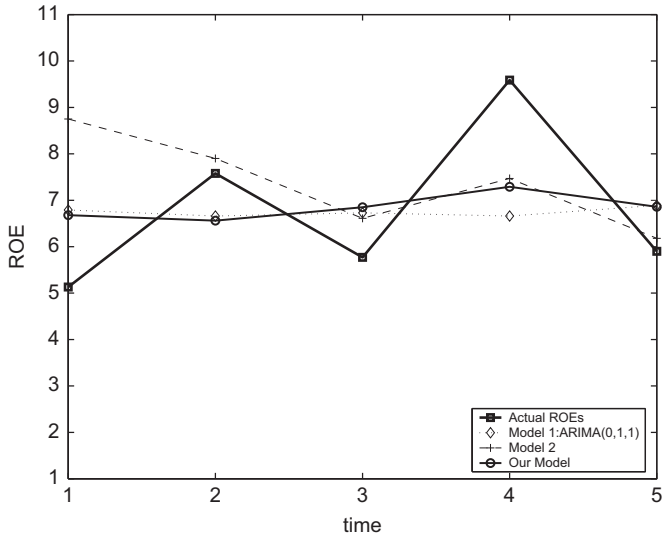

Fig. 3. Actual and predicted ROEs of SUNNING APPLIANCE when $N=20$.

3.2. Performance criteria

The prediction performance is evaluated using the following metrics, namely, MAE (mean absolute error), MAPE (mean absolute percent error), and RMSE (root mean square error), the definitions of these criteria are shown as follows:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t|, \tag{7}$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}, \tag{9}$$

where N is the number of forecasting periods, y_t is the actual ROE at period t, and $\hat{y}_t$ is the forecasting ROE at period t. The smaller the values of them, the closer are the predicted time series values to the actual values, that is to say, a smaller value suggests a better predictor.

3.3. Results on merchants bank

From the texts of Merchants Bank, 1644 terms are selected. Let the number of training set N be 20, 24, 28, and then the number of testing set will be 12, 8, 4. The appropriate parameters of different model are shown in Table 2.

Table 3(a)–(c) compares the forecasting results of the different models. Compiling all of the model performances together, our model performed best in all three metrics of all training sets. Those results indicate that the proposed model outperforms the other two models.

The actual and predicted values of the models are presented at Fig. 1(a)–(c). The point-to-point comparisons of results reveal that the proposed model yield better forecasting results than the other models.

3.4. Results on VANKE

From the texts of VANKE, 1900 terms are selected. Let the number of training set N be 20, 24, 28, and then the number of testing set will be 12, 8, 4. The appropriate parameters of different model are shown in Table 4.

Table 5(a)–(c) compares the forecasting results of different models. The actual and predicted values are plotted in Fig. 2(a)–(c). Compiling all of the model performances together, our model performed best in all three metrics except that the RMSE of Model 1 is slightly better than our method in the case that N is 20. Those results indicate that the model outperforms the other two models, revealing that our model is better able to capture ROE value movements.

3.5. Results on SUNNING APPLIANCE

From the texts of SUNNING APPLIANCE, 966 terms are selected. Let the number of training set N be 20, and then the number of testing set will be 5. The appropriate parameters of different model are shown in Table 6. Table 7 compares the forecasting results of different models. The actual and predicted values of the four models are presented at Fig. 3. Compiling all of the model performances together, our market sentiment based model performed best in all three metrics. Those results indicate that the proposed model outperforms the other two models.

3.6. Results on GE

From the texts of GE, 1742 terms are selected. Let the number of training set N be 44, 60, and then the number of testing set will be 24, 8. The appropriate parameters of different model are shown in Table 8.

Table 9(a)–(b) compares the forecasting results of different models. The actual and predicted values are plotted in Fig. 4(a)–(b). Compiling all of the model performances together, our model performed best in all three metrics except that the MAE and MAPE of Model 2 are slightly better than our method in the case that N is 44.

3.7. Results on Johnson&Johnson

From the texts of Johnson&Johnson, 1046 terms are selected. Let the number of training set N be 44, 60, and then the number of testing set will be 24, 8. The appropriate parameters of different model are shown in Table 10. Table 11(a) and (b) compares the forecasting results of different models. The actual and predicted values are plotted in Fig. 5(a) and (b). Compiling all of the model performances together, our model performed best in all three metrics. Those results indicate that the model outperforms the other two models, revealing that our model is better able to capture ROE value movements.

3.8. Results on McDonald's

From the texts of McDonald's, 1218 terms are selected. Let the number of training set N be 44, 60, and then the number of testing set will be 24, 8. The appropriate parameters of different model are shown in Table 12. Table 13(a) and (b) compares the forecasting results of different models. The actual and predicted values are plotted in Fig. 6(a) and (b). Compiling all of the model

performances together, our model performed best in all three metrics except that the MAE and MAPE of Model 2 are slightly better than our method in the case that N is 60.

Table 8
Parameters of different models on GE.

N	Model 1	Model 2	Our model
44	ARIMA (3,1,0)	$c=0.5$, $\gamma=256$, $\varepsilon=0.25$	$c=16.0$, $\gamma=0.0009765625$, $\varepsilon=0.0009765625$
60	ARIMA (3,0,0)	$c=8.0$, $\gamma=0.00390625$, $\varepsilon=0.5$	$c=4.0$, $\gamma=0.0009765625$, $\varepsilon=0.001953125$

Table 9
Performance comparison on GE when (a) $N=44$, (b) $N=60$.

Metrics	Model 1	Model 2	Our model
(a)			
MAE	0.66	0.61	0.65
MAPE	18.70	16.95	18.24
RMSE	0.87	0.87	0.86
(b)			
MAE	0.85	0.95	0.36
MAPE	34.76	38.98	13.31
RMSE	0.94	1.06	0.46

3.9. Overall results

To highlight how much gain one can get by using additional textual information in financial time series modeling and forecasting, the comparison of three models' overall performance is presented here. The average MAE, MAPE and RMSE of all performance listed above are shown in Table 14. These results indicate that our model outperforms the other two models. Compared with Model 1, out model obtain 12.41%, 12.82% and 10.98% drop rate for MAE, MAPE and RMSE. Compared with Model 2, out model's MAE, MAPE and RMSE drop by 20.43%, 25.10% and 17.67% respectively.

4. Conclusions

For more than half a century, the ARIMA model has been one of the most widely used linear models in the time series forecasting research area, but it is not adequate to capture the nonlinear patterns and it also requires large amounts of historical data. Recently, SVMs have achieved quite good performance in solving this problem. Hybrid models, which combine different models, have shown improved forecasting accuracy. But most of the existing forecasting models do not take the other factors, such as market sentiment, into consideration. To overcome these limitations, motivated by evidence that market sentiment contains some forecasting information, this paper has proposed a novel text mining approach via combining the ARIMA and the

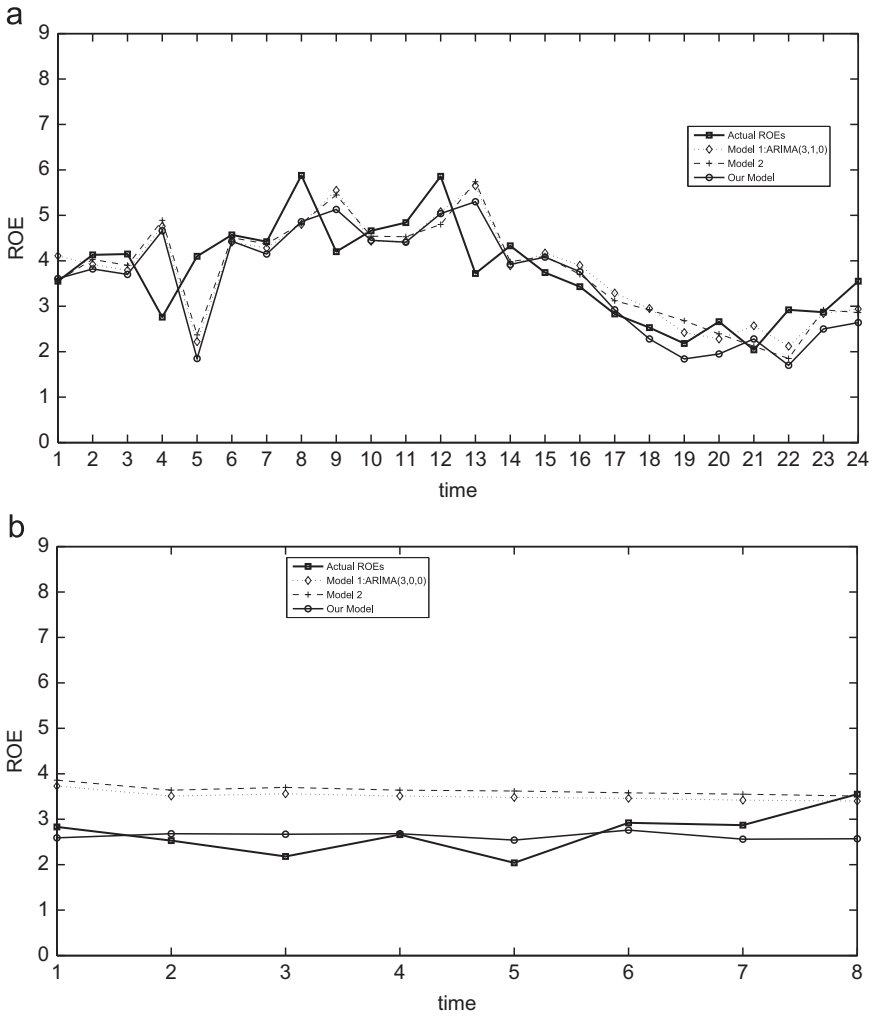

Fig. 4. Actual and predicted ROEs of GE when (a) $N=44$, (b) $N=60$.

SVR. For complex time series forecasting, such as financial time series forecasting, the market sentiment based method can be a promising way to improve forecasting accuracy. Hence, the novelty of the proposed method is that our model exploits the unique feature and strength of market sentiment in determining nonlinear patterns.

Table 10
Parameters of different models on Johnson&Johnson.

N	Model 1	Model 2	Our model
44	ARIMA (3,2,0)	$c=2.0$, $\gamma=64$, $\varepsilon=1.0$	$c=16.0$, $\gamma=0.0009765625$, $\varepsilon=0.125$
60	ARIMA (4,0,0)	$c=0.001953125$, $\gamma=512.0$, $\varepsilon=0.015625$	$c=2.0$, $\gamma=0.0009765625$, $\varepsilon=0.125$

Table 11
Performance comparison on Johnson&Johnson when (a) $N=44$, (b) $N=60$.

Metrics	Model 1	Model 2	Our model
(a)			
MAE	2.14	2.81	1.91
MAPE	31.0	41.8	29.1
RMSE	2.81	3.20	2.57
(b)			
MAE	1.05	1.08	0.99
MAPE	22.89	23.45	22.24
RMSE	1.45	1.47	1.42

Table 12
Parameters of different models on McDonald's.

N	Model 1	Model 2	Our model
44	ARIMA (4,0,0)	$c=1.0$, $\gamma=1024.0$, $\varepsilon=0.560924$	$c=4.0$, $\gamma=0.0009765625$, $\varepsilon=0.0009765625$
60	ARIMA (3,1,0)	$c=0.5$, $\gamma=0.5$, $\varepsilon=0.015625$	$c=2.0$, $\gamma=0.0009765625$, $\varepsilon=0.25$

Table 13
Performance comparison on McDonald's when (a) $N=44$, (b) $N=60$.

Metrics	Model 1	Model 2	Our model
(a)			
MAE	2.60	2.58	2.44
MAPE	32.75	32.37	30.94
RMSE	3.16	3.18	2.98
(b)			
MAE	0.72	0.67	0.69
MAPE	8.04	7.56	7.94
RMSE	0.86	0.83	0.83

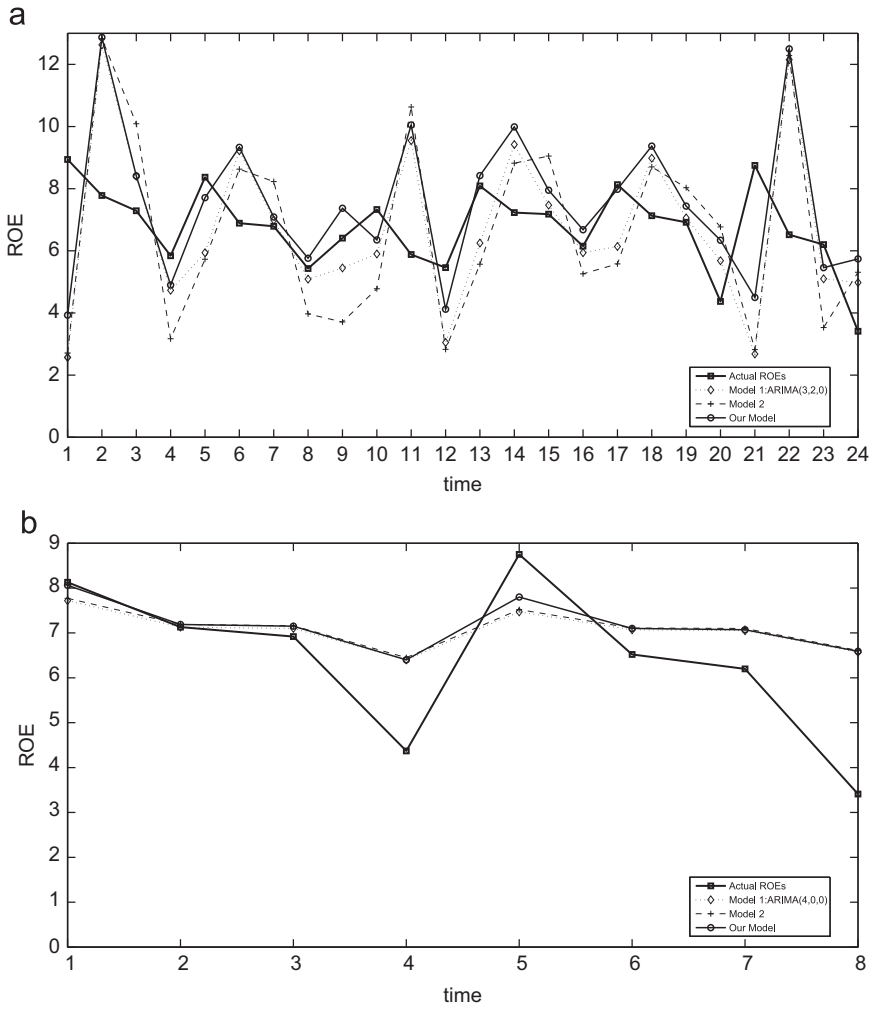

Fig. 5. Actual and predicted ROEs of Johnson&Johnson when (a) $N=44$, (b) $N=60$.

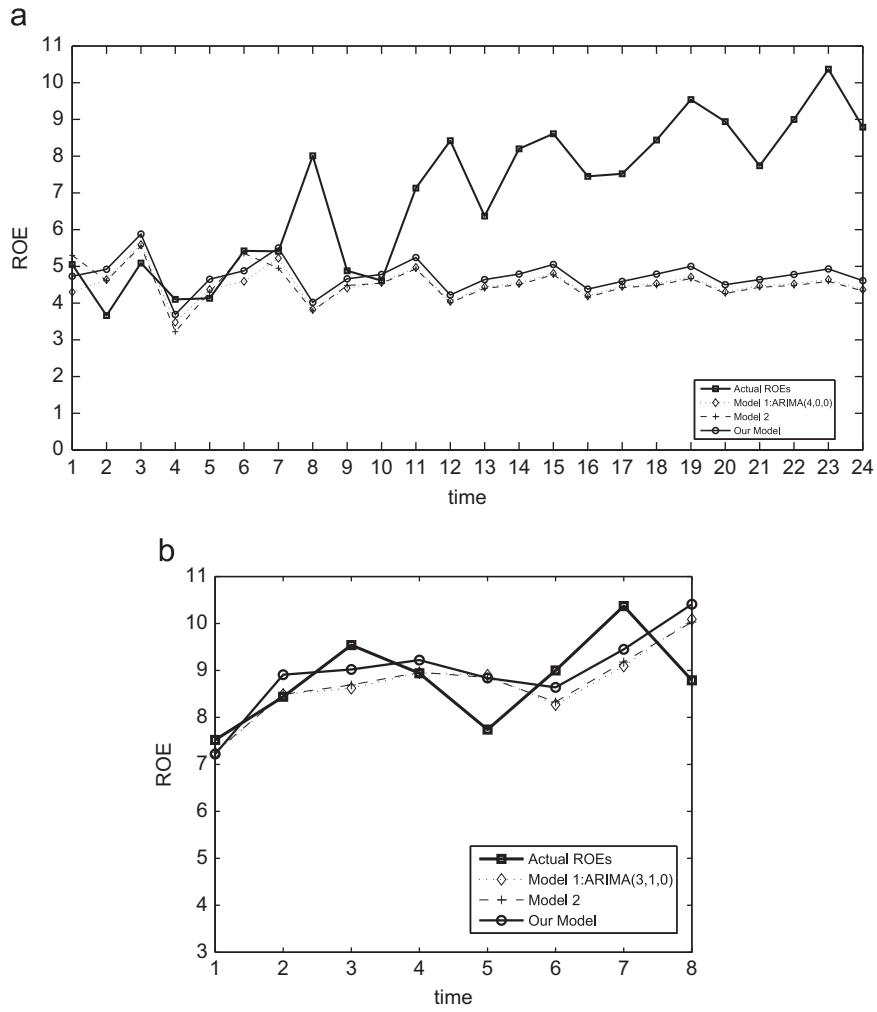

Fig. 6. Actual and predicted ROEs of McDonald's when (a) $N=44$, (b) $N=60$.

Table 14
Overall performance comparison.

Metrics	Model 1	Model 2	Our model
MAE	1.22	1.34	1.07
MAPE	26.45	30.79	23.06
RMSE	1.51	1.63	1.34

The proposed model has been evaluated with quarterly ROE time series of six companies, namely Merchants Bank, VANKE, and SUNING APPLIANCE from China A-share market and GE, Johnson&Johnson and McDonald's from US stock market. Experimental results show that the proposed model improves the prediction performance comparing with the single ARIMA model or hybrid model using only numbers. It indicates that the market sentiment based model provides a promising alternative to solve financial time series forecasting problem.

Finally, there are some caveats to be expressed. Market sentiment and time series data are often not in the same sampling rate, market sentiment may be missing at various time steps and some market sentiments may not be relevant to the time series data. Only those sentiments which have the same sampling rate with time series data and can truly reflect the values of the time series data can be selected as the forecasting sources. Furthermore, while the findings in this paper are interesting, we acknowledge that they rely on a small dataset, we plan to use a larger dataset in the next phase.

It is also a simplified approach to regard a nonlinear time series as a linear part and a nonlinear part. In a more general approach, a nonlinear model can be considered as a mixture of various linear models [22]. Since textual information can aid in all parts of the nonlinear hybrid model, we believe it can play more important role and make more contributions to prediction accuracy.

Acknowledgments

We would like to express our sincere appreciation to the anonymous reviewers for their insightful comments, which have greatly aided us in improving the quality of the paper. This work is supported by the High Technology Research and Development Program of China (2006AA01Z197), the Major Program of National Natural Science Foundation of China (No. 60435020), the National Natural Science Foundation of China (Nos. 60873168, 60603028, 61173075, 60973076, 60703015 and 61075037) and Shenzhen Municipal Science and Technology Plan (JC200903130224A, JC201005280522A).

References

[1] Y. Abu-Mostafa, A. Atiya, Introduction to financial forecasting, Appl. Intell. 6 (3) (1996) 205–213.

[2] G. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, Prentice-Hall PTR, Upper Saddle River, NJ, USA, 1994.
[3] M. Khashei, M. Bijari, G. Raissi Ardali, Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs), Neurocomputing 72 (4-6) (2009) 956–967.
[4] P. Pai, C. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega 33 (6) (2005) 497–505.
[5] K. Chen, C. Wang, A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, Expert Syst. Appl. 32 (1) (2007) 254–264.
[6] F. Palm, A. Zellner, To combine or not to combine? Issues of combining forecasts, J. Forecast. 11 (8) (1992) 687–701.
[7] D. Wedding, et al., Time series forecasting by combining RBF networks, certainty factors, and the Box–Jenkins model, Neurocomputing 10 (2) (1996) 149–168.
[8] H. Ni, H. Yin, Self-organising mixture autoregressive model for non-stationary time series modelling, Int. J. Neural Syst. 18 (6) (2008) 469–480.
[9] N. Terui, T. Kariya, Testing Gaussianity and linearity of Japanese stock returns, Finan. Eng. Jpn. Markets 4 (3) (1997) 203–232.
[10] N. Terui, H. Van Dijk, Combined forecasts from linear and nonlinear time series models, Int. J. Forecast. 18 (3) (2002) 421–438.
[11] J. Luxhoj, J. Riis, B. Stensballe, A hybrid econometric—neural network modeling approach for sales forecasting, Int. J. Product. Econ. 43 (2-3) (1996) 175–192.
[12] G. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.
[13] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 2000.
[14] V. Vapnik, S. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, in: Advances in Neural Information Processing Systems 9, Citeseer, 1996.
[15] F. Tay, L. Cao, Modified support vector machines in financial time series forecasting, Neurocomputing 48 (1-4) (2002) 847–861.
[16] K. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (1-2) (2003) 307–319.
[17] L. Cao, F. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Trans. Neural Networks 14 (6) (2004) 1506–1518.
[18] C. Chatfield, What is the best method of forecasting? J. Appl. Statist. 15 (1) (1988) 19–38.
[19] C. Lu, T. Lee, C. Chiu, Financial time series forecasting using independent component analysis and support vector regression, Decis. Support Syst. 47 (2) (2009) 115–125.
[20] F. Tay, L. Cao, Application of support vector machines in financial time series forecasting, Omega 29 (4) (2001) 309–317.
[21] S. Mukherjee, E. Osuna, F. Girosi, Nonlinear prediction of chaotic time series using support vector machines, in: Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop, IEEE, 2002, pp. 511–520.
[22] H. Ni, H. Yin, Exchange rate prediction using hybrid neural networks and trading indicators, Neurocomputing 72 (13–15) (2009) 2815–2823.
[23] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, A. Visa, Combining data and text mining techniques for analysing financial reports, Intell. Syst. Account. Finance Manage. 12 (1) (2004) 29–41.
[24] G. Gidófalvi, C. Elkan, Using News Articles to Predict Stock Price Movements, Department of Computer Science and Engineering, University of California, San Diego, 2001.
[25] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual web data, IEEE Int. Conf. Syst. Man Cybernet. 1998 3 (2002) 2720–2725.
[26] G. Pui Cheong Fung, J. Xu Yu, W. Lam, Stock prediction: integrating text mining approach using real-time news, in: 2003 IEEE International Conference on Computational Intelligence for Financial Engineering: Proceedings, 2003, pp. 395–402.
[27] R. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Trans. Inf. Syst. (TOIS) 27 (2) (2009) 1–19.
[28] S. Scott, S. Matwin, Feature engineering for text classification, in: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE, Citeseer, 1999, pp. 379–388.
[29] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Mach. Learn.: ECML 98 (1998) 137–142.
[30] G. Salton, Automatic text processing: the transformation, analysis, and retrieval of information by computer, in: Addison-Wesley Series in Computer Science, 1989, pp. 530.
[31] H. Zhang, H. Yu, D. Xiong, Q. Liu, HHMM-based Chinese lexical analyzer ICTCLAS, in: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, vol. 17, Association for Computational Linguistics, 2003, pp. 184–187.
[32] I. Feinerer, K. Hornik, D. Meyer, Text mining infrastructure in r, J. Statist. Software 25 (5) (2008) 1–54.
[33] C. Chang, C. Lin, LIBSVM: A Library for Support Vector Machines, available at 〈http://www.csie.ntu.edu.tw/cjlin/libsvm〉 (2001).

Baohua Wang is currently a PhD student in School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, and also a lecturer in the College of Mathematics and Computational Science, Shenzhen University. His research and teaching interests are in the area of natural language processing, machine learning and time series forecasting.

Hejiao Huang received the BS and MS degrees in Applied Mathematics from Shaan Xi Normal University in 1996 and 1999 respectively. Further, she received PhD degree in Computer Science from City University of Hong Kong in 2004. Her research interests include Petri net theory and applications, graph theory and application, optical and wireless mesh networks, machine learning and formal methods for system design. Huang is currently a Professor in the Department of Computer Science of Harbin Institute of Technology Shenzhen Graduate School, China.

Xiaolong Wang received the B.E. degree in Computer Science from the Harbin Institute of Electrical Technology, China, the M.E. degree in Computer Architecture from Tianjin University, China, and the Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology in 1982, 1984, and 1989, respectively. He joined Harbin Institute of Technology as an Assistant Lecturer in 1984 and became an Associate Professor in 1990. He was a Senior Research Fellow in the Department of Computing, Hong Kong Polytechnic University from 1998 to 2000. Currently, he is a Professor of Computer Science at Harbin Institute of Technology. His research interest includes artificial intelligence, machine learning, computational linguistics, and Chinese information processing.