Stock Price Forecasting by Combining News Mining and Time Series Analysis

Xiangyu Tang, Chunyu Yang, Jie Zhou
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Automation, Tsinghua University, Beijing, 10084, China
tangxy03@mails.thu.edu.cn, yangchunyu@mails.thu.edu.cn, jzhou@tsinghua.edu.cn

Abstract

Stock price forecasting has aroused great concern in research of economy, machine learning and other fields. Time series analysis methods are usually utilized to deal with this task. In this paper, we propose to combine news mining and time series analysis to forecast inter-day stock prices. News reports are automatically analyzed with text mining techniques, and then the mining results are used to improve the accuracy of time series analysis algorithms. The experimental result on a half year Chinese stock market data indicates that the proposed algorithm can help to improve the performance of normal time series analysis in stock price forecasting significantly. Moreover, the proposed algorithm also performs well in stock price trend forecasting.

1. Introduction

There is a huge variety of ways in forecasting economy indices like stock prices [1], most of which are time series analysis (TSA) methods based on structured data (e.g. stock prices table) [2], [3]. But since there are numerous factors influencing stock prices, and the relationship between those factors is also complicated, it hard to estimate the stock prices precisely just using the information of stock price itself, and the accuracy of those methods is limited.

We notice that economy news contains lots of information which would affect economic activities greatly. And with the development of communication, there are thousands pieces of economy news reports released every day, which we can easily access. Those news reports often reveal unexpected information and have a high influence factor with stock prices. Therefore, mining these news is of great value. But owing to the difficulty in relevant information extraction from these unstructured data, information in those data is rarely used in stock price forecasting. Some recent researches have shown that unstructured text data can be used to help forecasting stock prices' trend [4], [5], [6], which is a much simpler problem than price forecasting.

In this paper, we propose an algorithm to combine text mining techniques and time series analysis algorithms together for economy indices' forecasting. News reports can be automatically downloaded, categorized and analyzed, then the stock prices' changing can be forecasted based on both time series analysis methods and the news analysis results. We refer it NTF (News mining and Time series analysis based Forecasting) for short. In NTF, text mining is adapted to extract information related to stock prices' changing. Firstly, we perform normal TSA algorithm on stock prices data and obtain their forecasting results, then a regression-function is trained to quantify the information extracted in news mining and represented by chosen features, and to indicate how should those quantified information be weighted and added to TSA forecasting results for accuracy improving. Difference between stock data and improved result is minimized in training process. Then we use the regression-function to improve future normal TSA forecasts with information extracted in incoming news. NTF was tested on economy news and stock price data in Chinese Shanghai stock market from January 2008 to November 2008. The results indicate that NTF can provide forecasting results for stock prices and their trends, and they significantly outperform the forecasting results produced using previously proposed algorithms.

NTF differs from previous studies in that it integrates news mining with conventional stock price forecasting algorithms, and both stock prices value and trend forecasting can be performed by it.

2. Related Works

2.1. Text Mining Techniques

Feature extraction and feature selection are crucial steps in text mining. In feature extraction of Chinese documents, word segmentation proved to be necessary and difficult [7]. Fortunately, most of phrases in Chinese language contain only two characters, so bi-gram is an alternative way.

Many metrics are proposed for feature selection in text mining [8], and some of them have proved effective in many cases [9]. Probability Ratio is proved to work well in many text classifying problems [8]. After feature selection, we can simply represent our document collection by each feature's metric in each document.

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.48

2.2. Time Series Analysis Forecasting

Time series analysis is a large branch of forecasting methods. TSA algorithms and their improved version proved effective and efficient in many cases [10], [11].

In our work and many cases [1], stock prices have neither shown a clear periodic rule nor reflected very close to an ARMA model, and few algorithms perform well on many stock prices' data set. In this case, simple methods like MA rather than those complicated methods like ARMA, have better performances, which are measured by metric like sum of absolute differences rate.

Several works use text mining to help forecasting the trends of stock prices or indices [4], [6], and random walk model is also proved effective in this task [12]. They mainly categorize news releases, using the categorization results of news in earlier periods to forecast stock prices' trends in later periods. Our work extended their work from stock prices' trends to their values, we integrated time series analysis with text mining, and enhanced performance of predictor in stock prices values comparing to normal TSA. Simultaneously, we found performance of predictor in stock prices trends also enhanced comparing to random walk model.

3. The proposed algorithm

3.1. TSA module

One single stock price or index of one trade is thought to be a sequence varying with time in TSA. As mentioned above, since there are so many factors affecting stock prices, it is not realistic to estimate every factor exactly just using the price itself. So despite its simplicity, moving average (MA) algorithm is one of the most efficient and steadiest algorithms in stock prices' forecasting. It calculates average price of last several periods as a forecast for the next period, as shown in Eq. 1.

$$y_i^n = \sum_{j=n-T}^{n-1} \beta_{j-n+T} x_i^j / T, \tag{1}$$

where y_i^n means the n-th period forecasting of the i-th stock's price; x_i^j means the j-th period' stock price of the i-th stock.

3.2. News Analyzing and Supervised Learning

3.2.1. News Analyzing. Before preprocessing, repetitive news data are eliminated so as to perform text mining algorithms. As word segmentation and relevant algorithms [7] are adapted to solve the Chinese words segmentation problem, we process feature extraction and selection in NTF by following steps:

1) Manually create a list of words, which are most frequently used and related to the stocks we care.
2) Use the initial list to find relevant news document by finding whether the document contains words of the list, and features that occur both in the earlier half and later half are extracted.
3) Calculate PR table of all the selected features.
4) Create a blacklist of features which contains words empirically meaningless like auxiliary words.
5) Rank the features by PR descending order, and eliminate those features in the blacklist. Finally, choose the top N features(words) to be the representation vector.

Thus the meaningful features related with stock prices are gained, and then news documents collection can be represented by numeric model, which is a $m \times n$ feature matrix, where there are m documents in the collection , and n features are chosen.

Algorithm 1 Forecasting Algorithm

```
Initialize: Get $x_i^n, n = 1, 2, ..., N$ and $x_{ti}^n, n = 1, 2, ..., N_{train}$.

Train:
Calculate TSA forecasting results for stock prices on training data as following:
$\hat{y}_{ti}^{j+T_y} = \sum_{k=j}^{j+T_y-1} \beta_{k-j} y_{ti}^k / T_y, j = 0, ..., N_{train} - T_y$

Calculate linear compositions for representative vector on training data:
$\hat{x}_{ti}^{j+T_x} = \sum_{k=j}^{j+T_x-1} \theta_{k-j} x_{ti}^k / T_x, j = 0, ..., N_{train} - T_x$

Minimize forecasting error on training data
$E = \sum_{j=1}^{N_{train}} (y_{ti}^j - \hat{y}_{ti}^j - \alpha \hat{x}_{ti}^j)^2, j = 0, 1, ..., N_{train}$
with proper $\alpha, \beta, \theta, T_y$ and $T_x$.

Forecast:
for $j = 1$ to $N$ do
    $\hat{y}_i^{j+T_y} = \sum_{k=j}^{j+T_y-1} \beta_{k-j} y_i^k / T_y$        (2)
    $\hat{x}_i^{j+T_x} = \sum_{k=j}^{j+T_x-1} \theta_{k-j} x_i^k / T_x$        (3)
    $\bar{y}_i^j = \hat{y}_i^j + \alpha \hat{x}_i^j$        (4)
end for

Result: $\bar{y}_i$.
```

3.2.2. Supervised Learning. In this module, we firstly use SVR [13] to gain the weights of all chosen feature.

As mentioned above, news documents collection can be represented as a $m \times n$ matrix M, and each row vector of M is a representation of news for a certain day. Thus M can be the training matrix of the SVR and the vector composed by stock prices's changing of each day can be the relevant training vector. Then weight vector W is obtained by training SVR, and a vector of modifying value V^m is obtained by calculating MW.

From the ways we choose the features and MA forecasts, it is natural that we choose a linear composition of the modifying vector and add it to the forecasting result with certain weight. That is to say, moving average result of the modifying vector is added to MA result of the stock prices.

Let Y_i represents a stock prices vector at the i-th period. Assume that $x_i^n, n = 1, 2, \cdots, N$ is the representation of the 1 to N-th periods' news influence about the i-th stock's price, then we have Algorithm 1.

In the training step, binary search method is used to find the proper parameters for optimizing the forecasting result on training data (minimizing the forecasting error).

4. Experiments

4.1. Data Set Description

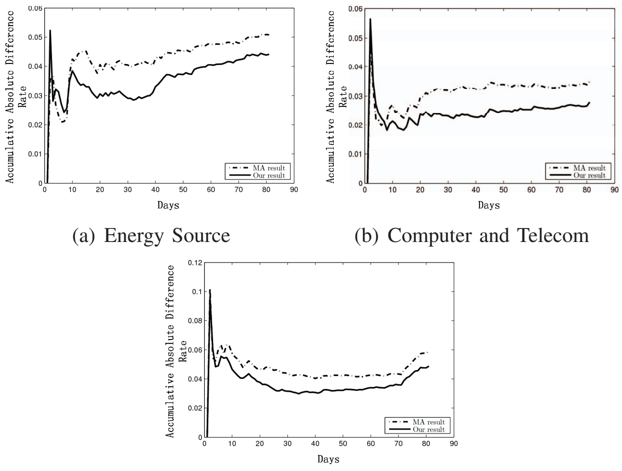
News articles used in our evaluation are from the RSS feeds of several large portals like http://news.baidu.com. They are collected continuously during March 2008 to July 2008 as crude news data. News titles and descriptions are chosen for text mining, their timestamps are natural tags to mark the position where they should be in time axis.

We use daily stock prices of the 841 A-type stocks on Shanghai Stock Market, and we use stocks' closing prices to stand for their prices of the day for convenience. In order to avoid sparseness in related news data, we choose three typical trades' indices to be our forecasting destination – energy sources, computer & telecom and real estate. And each trade is composed by a bundle of its related stocks, whose average price is thought to be the figure stands for the index of relevant trade.

4.2. Comparison

As moving average is a classical algorithm in time series analysis, and for its steady and efficient performance in stock price value forecasting [2], we compare our algorithm with it both in stock price value and trend forecasting. And since random walk is one of the best models for stock price trend forecasting [12], [6], it is compared with our algorithm in stock price trend forecasting in our experiment.

Average absolute difference rate is used to measure the accuracy of value-forecasting algorithms, noted by d. Accumulated absolute difference rate (d_{acu}) is calculated in every step to show detail process of the difference rate varying in periods: $d = \frac{\sum_{j=1}^{N} abs(y_i^j - \hat{y}_i^j)}{\sum_{j=1}^{N} y_i^j}$, $d_{acu} = \frac{\sum_{j=1}^{n} abs(y_i^j - \hat{y}_i^j)}{\sum_{j=1}^{n} y_i^j}$, where $n = 1, 2, ..., N$, and y_i^j is defined as the index of the i-th trade in the j-th period on test data, $\hat{y}_i^j$ is its forecast.



(a) Energy Source

(b) Computer and Telecom

(c) Real Estate

Figure 1. Forecasting difference rates of algorithms.

Similarly, average error rate and accumulated error rate are used to measure the accuracy of trend-forecasting algorithms, noted by e and e_{acu}, respectively. e is calculated as number of all error periods divided by number of all periods, e_{acu} is calculated as number of error periods from start to current divided by number of all periods.

4.2.1. Stock Price Forecasting. In our experiments, we compared the forecasting result of the neat MA algorithm in TSA and our integrated algorithm.

At first, we use the earlier half of the trade indices to train both algorithm's parameters, we search the parameter spaces to find the optimal ones, which minimize d on the training data. Then we use the trained parameter's to test on the later half of the trade indices data by calculating the d between each forecasting result and the original indices. Equation 1, 2, 3, 4 have shown how forecasting indices are calculated.

It is shown in Figure 1 that how accumulated absolute difference rates vary with the iterative calculation in the forecasting process of three trade indices. And d of both our method and MA in the trades are listed in Table 1.

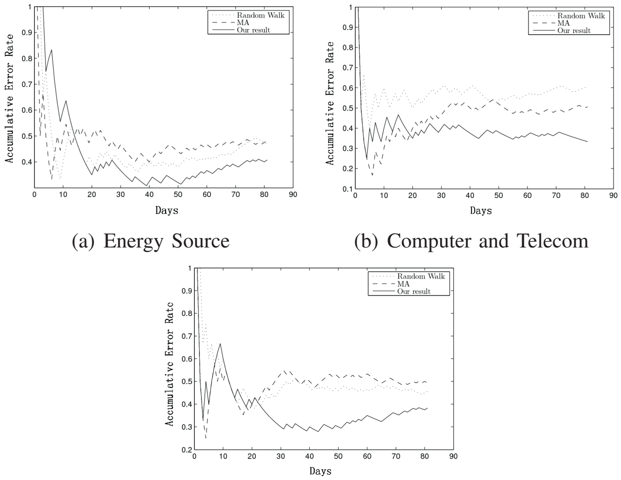
We can see that the accuracy of our forecasting is about 15% higher that accuracy of neat TSA algorithm in all three trades. So It can be seen that NTF algorithm outperforms MA in this trades' index's (stock prices') forecasting task in all trades in our experiment, and we can infer that extra information extracted from the news data can be used to improve the stock prices' forecasting.

4.2.2. Stock Price Trend Forecasting. Our trend forecasting result is directly derived from the value forecasting

Table 1. Average absolute difference rate

Trade	NTF	MA
energy sources	0.0441	0.0508
com & telecom	0.0277	0.0350
real estate	0.0488	0.0590

result, so is MA's trend forecasting result. Then we had Figure 2 to shown comparison of our method with random walk and MA in trend forecasting. And the error rate of trade index's trend forecasting in this experiment are listed in Table 2. Parameters are set by the same means of above.



(a) Energy Source

(b) Computer and Telecom

(c) Real Estate

Figure 2. Index trend forecasting results of algorithms.

Table 2. Error rate of trade index's trend forecasting

Trade	NTF	Random walk	MA
Energy Sources	0.40	0.48	0.49
Com & Telecom	0.33	0.60	0.51
Real Estate	0.38	0.45	0.50
Average	0.37	0.51	0.50

Then we can see that our algorithm significantly outperforms random walk and MA in the trades' indices trend forecasting task as well. Thus we can inform that our algorithm has a high forecasting accuracy in both prices' value and trend, and idea of improving Time Series Analysis by text mining is proved efficient on stock price data of Shanghai's daily stock market and news on the Internet.

5. Conclusion

In this paper, we proposed an integrated algorithm for forecasting stock price and other economic indices by incorporating both the power of text mining and time series forecasting technology. Experimental results on Chinese stock data using Chinese financial news articles spanning over half a year have shown that the proposed NTF performs well in forecasting the stock price comparing to regular TSA result, and NTF also performs well on forecasting the stock price trend comparing to classical random walk algorithm. It has proved that news reports can provide additional information for stock price forecasting and it gives an approach of improving conventional forecasting techniques.

References

[1] J. Stock and M. Watson, "Sforecasting output and inflation: The role of asset prices," *Journal of Economic Literature*, vol. 41, pp. 788–829, March 2001.

[2] D. Marcek, "Stock price forecasting: Statistical, classical and fuzzy neural network approach," in *MDAI*, ser. Lecture Notes in Computer Science, V. Torra and Y. Narukawa, Eds., vol. 3131. Springer, 2004, pp. 41–48.

[3] *A hybrid ARIMA and support vector machines model in stock price forecasting*, vol. 33, no. 3, 2005.

[4] B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam, "Daily prediction of major stock indices from textual www data," in *KDD*, 1998, pp. 364–368.

[5] V. Cho and B. Wüthrich, "Combining forecasts from multiple textual data sources," in *PAKDD*, ser. Lecture Notes in Computer Science, N. Zhong and L. Zhou, Eds., vol. 1574. Springer, 1999, pp. 174–178.

[6] M.-A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *HICSS*, 2004.

[7] X.-W. Zhang, M. R. Lyu, and G. zhong Dai, "Extraction and segmentation of tables from chinese ink documents based on a matrix model," *Pattern Recognition*, vol. 40, no. 7, pp. 1855–1867, 2007.

[8] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[9] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization," in *ICML*, D. H. Fisher, Ed. Morgan Kaufmann, 1997, pp. 143–151.

[10] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, UK: Cambridge University Press, 1990.

[11] A. J. Bagnall and G. J. Janacek, "Clustering time series from arma models with clipped data," in *KDD*, W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, Eds. ACM, 2004, pp. 49–58.

[12] T. Sandholm, "Autoregressive time series forecasting of computational demand," *CoRR*, vol. abs/0711.2062, 2007.

[13] S. R. Gunn, "Support vector machines for classification and regression," *ISIS Technical Report*, May 1998.