Financial Market Prediction Based On Online Opinion Ensemble

Wei Xu  
School of Information, Renmin University of China, and Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, Beijing, P.R. China, weixu@ruc.edu.cn

Daoyuan Sun  
School of Information, Renmin University of China, Beijing, P.R. China, dysun2010@126.com

Zihe Meng  
School of Information, Renmin University of China, Beijing, P.R. China, meng_zihe@hotmail.com

Meiyun Zuo  
School of Information, Renmin University of China, and Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, Beijing, P.R. China, zuomy@ruc.edu.cn

Follow this and additional works at: http://aisel.aisnet.org/pacis2012

Recommended Citation
Xu, Wei; Sun, Daoyuan; Meng, Zihe; and Zuo, Meiyun, "Financial Market Prediction Based On Online Opinion Ensemble" (2012). PACIS 2012 Proceedings. Paper 137.
http://aisel.aisnet.org/pacis2012/137
FINANCIAL MARKET PREDICTION BASED ON ONLINE OPINION ENSEMBLE

Wei Xu, School of Information, Renmin University of China, and Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, Beijing, P.R. China, weixu@ruc.edu.cn

Daoyuan Sun, School of Information, Renmin University of China, Beijing, P.R. China, dysun2010@126.com

Zihe Meng, School of Information, Renmin University of China, Beijing, P.R. China, meng_zihe@hotmail.com

Meiyun Zuo, School of Information, Renmin University of China, and Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, Beijing, P.R. China, zuomy@ruc.edu.cn

Abstract

Financial market prediction is one of the most attractive research areas in financial data mining field because of its available data and potential profits. Financial market prediction has been paid much attention by many researchers and practitioners. Huge numbers of forecasting methods including judgmental methods and statistical methods have been proposed. In previous studies, judgmental methods mainly focus on expert judgments, which are waste of resources in some extent and may lead to lower accuracy, while statistical methods in terms of statistical models can achieve better performance, but when the unpredictable events appear, these models are ineffective and useless sometimes. In this paper, by incorporating online user’s opinions, two novel paradigms are proposed for financial market prediction, which may overcome the aforementioned shortcoming. In the first paradigm, the opinions of several selective online users are extracted, and then their opinions are integrated to forecast financial market. In the second one, a data mining model is constructed by combining user’s opinions and financial time-series data, and then the model is used for financial market prediction. Moreover, these paradigms are validated and compared using real financial market data. The empirical results show that our proposed paradigms are useful and feasible for financial market prediction, and furthermore, the combined model outperforms the traditional judgmental model. These findings imply that the proposed method is a promising alternative for financial market prediction.

Keywords: Web mining, opinion mining, judgmental prediction, ensemble, financial market
1 INTRODUCTION

Financial market prediction is an attractive research area in financial data mining due to its available data and potential profits. A lot of researchers and practitioners pay attention to financial market prediction, and huge numbers of prediction models including judgmental models and statistical models have been proposed. In the previous forecasting studies, judgmental prediction, which mainly depends on experts’ experiences or large organizations’ knowledge (Lawrence et al., 2006), has played an important role in prediction (Webby and O’Connor, 1996). For example, Blair et al. (2002) used an expert judgment model to forecast the resurgence of the U.S. economy in 2001. Similarly, Blair et al. (2010) forecasted the resurgence of the U.S. economy in 2010. Remus et al. (1995) elicited the relationship between information reliability and forecasting accuracy in judgmental prediction, and the research results showed that human judgment could integrate information on anticipated time series changes and get better results. Bolger and Onkal-Atay (2004) studied the effects of feedback on judgmental predictions, and the results showed that after feedback, forecasts can be significantly improved. Onkal et al. (2012) reviewed the important role played by judgmental prediction, and examined consensus forecasts from structured groups with and without role-playing. Furthermore, Wright and O’Connor (1996) summarized and compared various judgmental forecasting methods, while Wright and Rowe (2011) reviewed on different group judgmental forecasting methods by integrating extant knowledge.

However, judgemental prediction mainly focuses on expert judgments, so these methods are waste of resources and sometimes may lead to lower accuracy. To improve the forecasting performance, statistical methods are suggested for financial market prediction. Initially, time series models such as ARIMA is offered to forecast financial market. For example, Thomakos and Jr (2004) compared the forecasting performance of ARIMA with the ones of other commonly used methods. Furthermore, several econometric models are offered. For example, Hung (2009) applied a new GARCH model to financial market and the performance is significantly improved in comparing with traditional GARCH model. With the development of artificial intelligence, some data mining methods have been developed to forecast financial market. For example, Refenes and Holt (2001) forecasted financial market volatility using neural networks. The intelligent forecasting methods for stock price prediction were also summarized in the literature (Atsalakis and Valavanis, 2009). Meanwhile, statistical methods including time series models, econometric models and intelligence models were compared in the existing literatures. Kanas and Yannopoulos (2001) compared linear and nonlinear forecasting method for stock returns. Similarly, Cao et al. (2005) compared the predictive power using ANN models and econometric models, and the results showed that ANN models outperformed other forecasting models. Furthermore, some hybrid statistical approaches incorporating these techniques were proposed. Roh (2007) proposed hybrid neural network and GARCH family models to forecast the volatility of stock prices, while Pai and Lin (2005) hybridized ARIMA and SVM model. Similarly, Khashei et al. (2008) combined neural network model with fuzzy regression model for financial market prediction.

Although the statistical methods show high accuracy in financial market prediction, when the unpredictable events appear, these models are ineffective and useless sometimes. So, to combining the advantage of judgmental models and statistical models, some integrated forecasting methods have been developed and the examples in the existing literatures are presented. Sanders and Ritzman (1995) brought judgment to combine forecasts. Goodwin (2000) integrated statistical model and judgement to improve forecasting accuracy. Similarly, Goodwin (2002) combined management judgement with statistical methods for improving the forecasting performance. Moreover, Cheikhrouhou et al. (2011) presented a judgemental collaborative approach for demand forecasting. In their method, the mathematical forecasts were considered as the basis, and then were adjusted by the structured and combined knowledge from different forecasters.
Recently, web information has been introduced to forecast financial market. Initially, news articles are used in financial market prediction. Mittermayer (2004) employed text mining techniques to forecast financial market using news articles, while Chan and Franklin (2011) developed a text based decision support system. Similarly, Schumaker and Chen (2009) developed a quantitative financial prediction system based on financial news. Furthermore, user generated content (UGC) is also introduced to improve the forecasting performance of financial market. Das and Chen (2007) extracted investor opinions from stock message boards, and forecasted stock market using the opinions. Similarly, Sehgal and Song (2007) scanned for financial message boards and extracted opinions expressed by individual users for financial market prediction. Finally, Bollen and Mao (2011) used the opinions from Twitter for financial market prediction.

In the prior UGC based financial prediction studies, users’ opinions from web are firstly gathered and then the relationship between the gathered opinions and financial market is modelled. This paper investigates the benefits in forecasting performance by incorporating online user’s opinions. Our study differs from previous research efforts on two important dimensions. First, two novel paradigms are suggested by combining online user’s opinions instead of experts or large organizations. In the first paradigm, the opinions of several selective online users are extracted, and then their opinions are integrated to forecast financial market. In the second one, a data mining model is constructed by combining user’s opinions and financial time-series data, and then the model is used for financial market prediction. Second, our proposed forecasting method mainly depends on several selective online users, but not the sentiments in the whole communities. So, our proposed method is a new alternative for judgmental prediction and forecast ensemble.

The rest of this paper is organized as follows. The theoretical background is introduced in Section 2. Section 3 proposes an online opinion based ensemble method for financial market prediction. For validation purpose, the empirical analysis of the proposed paradigms in real financial market data is reported in Section 4. Finally, conclusions and future research work are summarized in Section 5.

2 THEORETICAL BACKGROUND

In this section, some useful theories including ensemble learning theory and neural networks are introduced, and these theories are critically important to construct our proposed method.

2.1 Ensemble Learning Theory

The basic idea of ensemble learning theory for prediction tasks is to use each predictor’s unique feature to capture different patterns, and then combine their outputs with a fusion strategy (Czyz et al. 2004). In the ensemble learning theory, individual predictor can be represented as follows:

\[ F_i = A + \varepsilon_i \]  

(1)

where \( F_i \) is forecast value, \( A \) is actual value, \( \varepsilon_i \) is forecast error, and \( i (i=1, \ldots, m) \) stand for predictor index. As individual predictors may be adjusted up or down for negative or positive errors, for the purpose of reducing the errors, a weighted average of the forecast values of individual predictors can be implemented for ensemble learning (Maines, 1996). The ensemble prediction can be represented as follows:

\[ F_e = \sum_{i=1}^{m} k_i F_i \]  

(2)

where \( F_e \) is the combined forecast value, \( k_i \) are weights for individual predictors, and \( \sum_{i=1}^{m} k_i = 1 \).
To improve the effectiveness and efficiency of ensemble learning, two aspects including how to construct diversified predictors and how to combine a set of individual predictors can be considered. In the first aspect, some methods, such as stacking (Breiman, 1996), bagging (Mao, 1998), boosting (Mao, 1998), and random subspace (Ho, 1998), are suggested for diversify the predictors. For example, stacking is to combine different predictors generated by various methods to improve prediction accuracy (Breiman, 1996). In the second aspect, some methods including simple average, outperformance (Bunn, 1975), and optimal (Bates and Granger, 1969) are proposed. For instance, outperformance method combines predictors into $F_e = \sum_{i=1}^{m} p_i F_i$, where $p = (p_1, ..., p_m)$ is a probability distribution which can be assessed in a Bayesian way (Menezes et al. 2000).

2.2 Introduction to Neural Networks

Neural networks (NNs), which attempt to simulate biological neural network, are composed of groups of artificial neurons, and use computational models to process various kind of information. In a NN, simple nodes (neurons) connect together to exhibit complex behavior, and it is able to relate inputs and outputs and recognize pattern in data (Wasserman, 1989). So, it is a useful tool for nonlinear modeling and prediction. One of the most common NN models is the back-propagation neural network (BPNN). It could be regarded as a multi-stage adaptive system which is trained to optimize the object function. The structure of a BPNN is illuminated in Figure 1.

![Figure 1. The structures of a neural network](image)

As can be seen from Figure 1, BPNN for a prediction task can be implemented in the following steps:

Step 1: Data collection. Research dataset is collected and selected according to our research target, and then the data set can be divided into training set and testing set.

Step 2: Training. In the training process, two phases including forward propagation and back propagation are carried out. In the forward propagation, it attempts to generate output activation, and then in the backward propagation, it generates deltas of all output and hidden nodes. Next, input activation and output deltas are multiplied to get the gradient of weight. BPNN transfers information from input to output layer while adjusting errors in the reversed process. Finally, the errors continue to change until the network output satisfies the error criteria.
Step 3: Testing. Through testing, the forecasting performance of NNs with different structures and various parameters are compared, and a proper NN with optimal structure and appropriate parameters is selected for a prediction task.

Step 4: Prediction. When the testing is finished, a BPNN with optimal structure and proper parameters is used to forecast the trend.

3 THE PROPOSED METHOD

In this section, an online opinion based ensemble method is proposed for financial market prediction. In our proposed method, several proper online users are firstly selected for the purpose of the prediction task. In the following, two paradigms based on online opinion ensemble are proposed to forecast financial market. In the first paradigm, the opinions of the selective online users are combined to forecast financial market. In the second one, a data mining model is constructed by combining user’s opinions and financial time-series data, and then the model is used for financial market prediction. Among these paradigms, the first one belongs to judgmental prediction, while the second one is an integrated model combined judgmental prediction and statistical prediction. Finally, these paradigms are validated and compared using real financial market data, and through model selection, the proper model is selected and used for financial market prediction. The framework of the proposed online opinion based ensemble method is illustrated in Figure 2, and the details will be discussed in the following subsections.

![Figure 2. The framework of the proposed method](image)

3.1 Online User Selection

Selection of proper online users is critically important for financial market prediction due to some reasons as follows. First of all, if an online user does not usually give opinions, judgmental information cannot be generated. So, the frequencies on giving opinions by online users in a certain time window cannot be low. Second, instead of forecasting experts’ opinions, the forecasting accuracy
of online users’ opinions cannot be low. Finally, the topics of online users’ opinions should relate to financial areas. So, according to the aforementioned criterion, online users can be selected. When the online users are fixed, tweets of these online users can be crawled by a spider.

3.2 The First Paradigm: Judgmental Prediction

When the proper online users are confirmed and tweets from Twitter have already been collected, judgmental prediction can be implemented in the following steps:

Step 1: Extracting online users’ opinions. During the process of valuing the opinions of tweets, OpinionFinder is firstly employed to classify the strength of a single word’s sentiment by using MPQA (Multi-Perspective Question Answering). Furthermore, the strength of every single tweet can be calculated according the words in a tweet. Finally, three daily time series of online users’ opinions including positive opinion ($P_t$), negative opinion ($N_t$) and compositive opinion ($U_t$) are calculated. $U_t$ can be calculated by $U_t = P_t - N_t$.

Step 2: Combining online users’ opinions. Suppose financial time series be $F_t$, the predictive value $F_{t+1}$ can be represented by the formula:

$$F_{t+1} = \sum_{i=1}^{m} k_i F_{it+1}$$

where $k_i$ are weights for user $i$, $\sum_{i=1}^{m} k_i = 1$ , and $F_{it+1}$ can be calculated by the formula

$$F_{it+1} = \alpha + \beta U_{it} + \epsilon$$

The weight $k_i$ is determined by the forecasting accuracy of user $i$. The better performance the user $i$ achieved, the more value the weight $k_i$ has.

Step 3: Forecasting financial market trend. When the weight $k_i$ is confirmed, the formula (3) can be used to forecast financial market trend.

3.3 The Second Paradigm: Integrated Prediction

When the proper online users are confirmed and tweets from Twitter have already been collected, integrated prediction which fuses judgmental model and statistical model can be implemented in the following steps:

Step 1: Extracting online users’ opinions. According to the ways mentioned in subsection 3.2, three daily time series of online users’ opinions including positive opinion ($P_t$), negative opinion ($N_t$) and compositive opinion ($U_t$) are calculated.

Step 2: Constructing neural network model. In order to construct neural network model, some of the opinion time series including $P_t$, $N_t$ and $U_t$ and financial time series $F_t$ and their lags are chosen as input of NN, and financial time series $F_{i+1}$ is used as output of NN. The details of the neural network modelling process for user $i$ are shown in Figure 3.

Step 3: Combining neural network models. Suppose the predictive value of neural network model for user $i$ be $F_{it+1}$, the predictive value $F_{t+1}$ can be represented by the formula:

$$F_{t+1} = \sum_{i=1}^{m} k_i F_{it+1}$$

where $k_i$ are weights for model $i$, and $\sum_{i=1}^{m} k_i = 1$. 
Similarly, the weight $k_i$ is determined by the performance of model $i$. The better performance the model $i$ achieved, the more value the weight $k_i$ has.

Step 4: Forecasting the financial market trend. When the weight $k_i$ is confirmed, the formula (4) can be used to forecast financial market trend.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (F_i - \hat{F}_i)^2}$$

where $F_i$ is the actual closing prices and $\hat{F}_i$ is the forecast ones, and $N$ is the number of data.

4.2 Experimental Results

Before the experiment are done, online user selection is first considered. In our experiment, the top 40 active users from Twitter are selected, and about 190132 tweets from 3rd August, 2009 to 30th November, 2009 are extracted. The OpinionFinder is employed to identify the sentiments of each user. After mining online users’ opinions, the time series of users’ opinions can be calculated, and three set of time series including $P_{it}$, $N_{it}$ and $U_{it}$ ($i = 1, ..., 40$) can be got. However, the time series may contain null values, indicating the online user has no tweets on that day, and null values may occur quite frequently. To better establish our paradigms, null values are filled in according to users’ performance.
4.2.1 Experimental Results: The First Paradigm

According to subsection 3.2, compositive opinion $U_i$ of user $i$ ($i = 1, \ldots, 40$) can be obtained, and data from 3rd August, 2009 to 30th October, 2009 is used to calculate the RMSE of each user. In terms of RMSE of each user, the weight of each user can be calculated by the following formula:

$$\text{Weight}_i = \frac{1}{\text{RMSE}_i} \frac{1}{\sum_{i=1}^{n} \frac{1}{\text{RMSE}_i}}$$

where $n$ stands for the number of users, which is set as 10, 20, 30 and 40 in our experiments. According to testing set, the experimental results are shown in Table 1.

<table>
<thead>
<tr>
<th>Number of users</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>826.77</td>
<td>802.67</td>
<td>813.25</td>
<td>817.38</td>
</tr>
</tbody>
</table>

Table 1. The RMSE of judgmental prediction

As can be seen from Table 1, the RMSE of the model within 20 users outperforms other models, and furthermore, the average RMSE of four models is less than the average one of each user. The experimental results reveal that the judgmental models have better performance than individual users.

4.2.2 Experimental Results: The Second Paradigm

According to subsection 3.3, there are three opinion time series including $P_i$, $N_i$ and $U_i$. In our experiments, the combination of different opinion time series are implemented, which is shown in Table 2.

<table>
<thead>
<tr>
<th>Group number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion time series</td>
<td>$P_i$</td>
<td>$N_i$</td>
<td>$U_i$</td>
<td>$P_i, N_i$</td>
<td>$P_i, U_i$</td>
<td>$N_i, U_i$</td>
<td>$P_i, N_i, U_i$</td>
</tr>
</tbody>
</table>

Table 2. Opinion time series used in our experiments

For each user, all these seven groups are tested by BPNN. The number of input nodes ($IN$) is equal to the number of opinions used in each group, and the relationship between the hidden nodes ($HN$) and the input nodes ($IN$) is followed by the experiential formula: $HN = \log_2(IN)$. In BPNN, the gradient descent with momentum and adaptive learning rate back propagation ($gdx$) is selected as the training function, and the learning rate is set as 0.05. Besides, the output of BPNN in testing set is used to calculate RMSE. The weight of each user is determined by the following formula:

$$\text{Weight}_i = \frac{1}{\text{RMSE}_i} \frac{1}{\sum_{i=1}^{n} \frac{1}{\text{RMSE}_i}}$$

where $n$ represents the number of users who are responsible for prediction, which is also set as 10, 20, 30 and 40. Therefore, 28 models are built and tested, and the RMSE of each model is shown in Figure 4 and Table 3.
As can be seen from the Table 3, the model with opinion time series $N_i$ and the 10 users outperform the other models. Therefore, time series information is added to this model and the RMSE is shown in Table 4.

<table>
<thead>
<tr>
<th>Time series</th>
<th>$N_i, F_{t-1}$</th>
<th>$N_i, N_{i-1}, F_{t-1}, F_{t-2}$</th>
<th>$N_i, N_{i-1}, N_{i-2}, F_{t-1}, F_{t-2}, F_{t-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>333.82</td>
<td>409.60</td>
<td>369.83</td>
</tr>
</tbody>
</table>

As can be seen from Table 4, by adding time series information, the forecasting performance can be obviously improved. Meanwhile, the RMSE of model with one lag of negative opinion and financial time series reaches lowest RMSE.
4.2.3 Comparison and Further Discussion

In this subsection, these models including judgmental prediction and integrated prediction are compared. As can be seen from Table 1, Table 3 and Table 5, the performance of integrated model with financial time series prevails over integrated model without financial time series and judgmental prediction. Moreover, as can be seen from Table 3, the model with negative opinions outperforms other integrated models. It implies that the negative opinion may critically affect financial market.

5 CONCLUSIONS

This paper presents a novel online opinion based ensemble method for financial market prediction. By incorporating online users, instead of forecasting experts, two paradigms including a judgmental prediction and an integrated prediction are proposed. In terms of evaluation criteria, the empirical results reveal that among the proposed models, the integrated models outperform judgmental prediction, and moreover, the integrated model with negative opinion and financial time series is dominant. Furthermore, the experimental results find that the negative opinion may accelerate financial market volatility. Meanwhile, the experimental results also indicate that our proposed method provides a new way for judgmental prediction, and the proposed online opinion based ensemble method is a potential and feasible alternative to mine financial market trend.

In addition, this study also has some research questions for further research. Firstly, since online user selection is an important issue for financial market prediction, how to evaluate online users objectively and effectively can be further considered, and moreover, the relationship between the number of online users and the forecasting performance can also be investigated. Secondly, since this paper suggests a novel way for judgmental and integrated forecast, some other paradigms may be developed to improve the forecasting performance. Thirdly, with the help of online users, a human-computer interaction system for financial market prediction may be developed. Finally, the proposed method may also be applied to other research areas, such as real estate market prediction, exchange rate forecasting and e-commerce fraud detection.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable suggestions which have helped greatly in improving the quality of this paper. This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No. 70971130, and 71001103), Beijing Natural Science Foundation (No. 9122013, and 9112009), the Program for New Century Excellent Talents in University, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China. The fourth author is the corresponding author.

References


