Forecasting Stock Market with Social Media Sentiment Based on Adaptive Network Fuzzy Inference System

Shen-lin JIANG and Wei-jie WU
Beijing University of Posts and Telecommunications, Beijing, 100876, P.R. China

Keywords: ANFIS, Social media, Stock market forecast.

Abstract. The Adaptive Neuro Fuzzy Inference System (ANFIS), based on fuzzy inference system, describes the process of human logical reasoning by establishing a number of fuzzy rules. It has a good ability to deal with uncertain and imprecise systems and is suitable for application in stock market forecasting. In this paper, the model inputs are chosen based on multiple feature combinations with multiple experiments, and social media is added as part of the features. Gaussian membership function was selected as the main constraint and triangle function was the result optimizer. The experiments using data from both social media and stock markets—the Sina Weibo and the Shanghai Stock Market—to train and evaluate the change in trend for the next trading day, with the “buy and hold” strategy and several other timing trading strategies. The empirical results show that our proposed method outperforms the ANFIS with only technical index as input.

Introduction

Stock market forecasting needs to be combined with a variety of computing techniques. Researchers have put forward some new ways to create new and better predictive results. In recent years, artificial neural networks (ANNs) and support vector machines (SVMs) have been successfully applied to solve the prediction of financial time series, including the financial stock market forecast. The neural fuzzy system is a good example.

Neural network pattern recognition and adaptation to the changing environment, and fuzzy reasoning systems will make the decision-making behaviors more reasonable. These two complementary approaches are integrated in the results of the neural fuzzy system model [1]. The use of intelligent systems, such as neural networks, fuzzy systems and genetic algorithms, has a wide range of applications in the financial field.

We already know that the volatility of the stock market depends on a variety of factors. With the trend of financial business online marketing and more ordinary people to participate in the financial markets, public opinion, whether from the forums or from micro-blog (Twitter), has an increasing impact on the trend of the stock market. Based on the above discussion, we believe that the use of social networks, to predict the future trend of the market, will provide favorable results for market forecasts.

To this end, we propose a new model to predict the trend of the stock market. First, we combine public opinion with technical indicators in the real market in order to adjust the overall direction of the forecast. In addition, we develop a public opinion model that will influence the user as a feature of our training in social media. The advantage of using influential users is that they can predict the general trend of the market; they enjoy great popularity among other users, and their comments and opinions on the stock market have high accuracy. Finally, we use ANFIS as an innovative technology to predict financial markets. In this paper, we use the technical indicators of the Chinese stock market, comments from SINA micro-blog (weibo.sina.com) and our proposed ANFIS model. The motivation for this article is to use historical data on stock prices to predict the challenges of the second day of the stock market trend.

The rest of the paper is presented in the following sequence. Section 2 provides a review of the prior literature. Section 3 describes the research design and experiments. Section 4 provides a detailed analysis of the experimental results. Section 5 discusses the conclusions and findings of the study.
Literature Review

The Application of ANFIS

Abraham et al. [2] introduced a kind of forecasting stock index for genetic programming technology, considering the NASDAQ stock market Nasdaq-100 index and S & P CNX NIFTY stock index as the test data. Then the performance and use of the Levenberg-Marquardt algorithm, support vector machine, Takagi-Sugeno neural fuzzy model and compare enhanced artificial neural network neural network training.

Cheng et al. [3] told us that ANFIS is used to study the effects of expected actions on the price movements and the volume changes of the U.S. Based on the expected impact of events on the market, investors will take advantage of the decision to carry out arbitrage in the market after the event is announced. The purpose of this study is to help investors make more informed decisions in this context.

The Application of Social Media Model

Researchers have never ceased to explore the relationship between public opinion in social media, and some progresses have been made so far.

Junqué de Fortuny et al. [4] proposed a new model based on the most advanced text mining techniques to predict the movement of stock prices and discuss the parameters suitable for different situations.

Smailovic et al. [5] presented a static Twitter data analysis problem in order to determine the best text preprocessing settings for training support vector machine (SVM) mood classifiers.

Yet exceptions also happen that some experiment exhibited an irrelevance between sentiment from public and market movement.

The Proposed Method

Outline

In this section, the outline of our method is shown as follow Figure 1.

As can be seen from Fig. 1, there are two main steps before the model training process. We will extract the keywords “stock market” from all the user’s comments in the social network, which will produce a lot of data. In accordance with the weighted relation in the social network, we will select in a few specific users and their views, as one of the eigenvalues of the ANFIS model, and will select the technical indexes of the multiple stock markets to combine these indicators.

Data Collection

We will get some user comments on Weibo.com. As one of the eigenvalues of the ANFIS model, and will select the technical indexes of the multiple stock markets to combine these indicators. Sina
Weibo is one of the largest social sharing platforms in China. It has a complex social relationship map and an open information environment. Public information published by anyone can be retrieved by keywords. Based on this environment, the user's information is released timely, other users can see the topic or keyword-related information at once, and the information can be real-time reflection of the user's emotional state. Therefore, the analysis of public opinion on this platform will be reasonable and persuasive.

**Sentiment Quantization**

As the procedure of data acquisition containing all the content related to the keyword, we need to remove the unrelated data before we carry out the calculation of the emotion value. Then we use the Chinese dictionary produced by the Chinese Academy of Sciences to segment the data and avoid the error when the dictionary matches the content of the article. We use the Chinese financial emotional dictionary and determine the emotional value of each word by using the relevant theory. According to this emotional dictionary, we calculated the emotional value of each user who posted information during the trading day.

Then we calculate the daily microblogging sentiment of each user by the following formula:

\[
Sent(i) = \left\{ \begin{array}{ll}
\sum_{\text{Lexicon}} Freq(word_j) \cdot Polarity(word_j), & \text{publish at least one microblog in day } i \\
Rand(i), & \text{no microblog published in day } i
\end{array} \right.
\]

In the formula, if a user publishes multiple messages in a day, we use the emotional values of all the words in the emotional dictionary to multiply their frequencies and sum the results.

If the user does not publish microblogging content on the day of the transaction, the corresponding emotional value will be lost. There are usually three missing values: Total Random Loss (MCAR), Random Loss (MAR), and Missing Random (MNAR).

The \( Rand(i) \) uses the ARIMA regression model to determine what determines the stock market index for a given day. ARIMA stimulus shows that the trend of AR (1) in the 2013-2014 have the trend of AR (1), i.e. \( X_t = \alpha_1 \cdot X_{t-1} + \varepsilon_t \), and the parameter \( \alpha_1 \) is 0.384. Therefore, we come up with the \( Rand(i) \) as:

\[
Rand(i) = \left\{ \begin{array}{ll}
\alpha_1 \cdot Sent(i - 1) + [\varphi_1 \cdot \text{random}(-1,1) - \omega_1 \cdot \text{random}(-1,1)], & \exists Sent(i - 1) \\
Sgn(diff(X_{t-1}))[\varphi_2 \cdot \text{random}(-1,1) - \omega_2 \cdot \text{random}(-1,1)], & \text{noSent}(i - 1)
\end{array} \right.
\]

The function indicates that the mood assumed in the day is determined by yesterday's mood and yesterday's trend of the stock market, since the former was generated by the simulation of the ARIMA model. The latter is due to the assumption that people's current opinions are usually influenced by yesterday's performance. If the stock market is working well yesterday, then a person is likely to still think it will flourish today. In the function, \( \varphi_1, \omega_1, \omega_2 \) are the parameters that measure the likelihood of past emotional recurrence. These parameters can be changed during the experiment to achieve the best results. Usually we will set \( \varphi_1 > \omega_i, i = 1, 2 \).

**Training Process with ANFIS**

Classically, training ANFIS controllers based on inverse learning techniques (Jang et al., 1997). The ANFIS controller can select multiple inputs to simultaneously select a single output to be used together to calculate the price (output) at the time of \( y(k) \) and input \( u(k) \) based on the general training data set [approximate inverse mapping G. \( y(k + 1) \)] At the previous stock price. After the training phase then, given the desired future stock price, the ANFIS controller will generate the estimated price. As more data sets are used to improve the parameters in the ANFIS controller, the prediction results will be closer to the real results, and as the training process continues, the control will be more accurate. The reason for using the Sugeno first-order model is because the ri parameter is used to approximate the better real value. The error back propagation gradient descent method is used to optimize the parameters of the first part (prerequisite) of the rule and the least squares error method is used to optimize the parameters of the second part (consequence).
Stock Trading Strategy

In the experiment, we will use the index as the result of the prediction, but to predict the index accurately is an almost impossible thing. As a result, we will transform the index of these different models into the prediction of Index fluctuation.

Through the empirical test results of different initial threshold analysis, we found that when the initial threshold was set as 1%, the effect is relatively good. We also set a stop trading line. If a single timing loss of more than 10% that remain empty positions, until the timing signal changes.

In order to get closer to the reality of the actual transaction, we will set the transaction fee for each operation 0.5%.

Experiments Analysis

Data Description and Evaluation Criteria

As the CSI 300 index can better reflect the trend of China's stock market, we chose the CSI 300 index for the analysis of the object. In order to analyze the trend performance of stable and volatility models, we selected 970 samples of daily data from January 3, 2010 to December 31, 2014.

In general, the selection of input variables has great influence on the forecasting effect of ANFIS model, and the factors that can best reflect the change of stock price fluctuation can improve the accuracy of forecasting. In this paper, 12 market indicators variable or technical indicator variables are chosen as the original data, see Table 1, in addition, in order to describe the trend of variable trends, we selected 12 variables lag 1, lag 2 and lag 3 as the original data derived variable to enter the model, so there is a total of 48 input variables.

Table 1. Indexes and variable description.

<table>
<thead>
<tr>
<th>Indexes</th>
<th>Variable description</th>
<th>Indexes</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turn volume</td>
<td>Index of the daily turn</td>
<td>MACD</td>
<td>MACD (26,12,9) MACD</td>
</tr>
<tr>
<td></td>
<td>volume</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close</td>
<td>Index of the daily close</td>
<td>EMA5</td>
<td>Closing price of EMA5 moving average</td>
</tr>
<tr>
<td>Change</td>
<td>Index of the daily change</td>
<td>K</td>
<td>(14)</td>
</tr>
<tr>
<td>High</td>
<td>Index of the daily highest</td>
<td>D</td>
<td>(14)</td>
</tr>
<tr>
<td>Low</td>
<td>Index of the daily lowest</td>
<td>J</td>
<td>(14)</td>
</tr>
<tr>
<td>DIFF</td>
<td>MACD (26,12,9)DIFF</td>
<td>RSI</td>
<td>RSI (14)</td>
</tr>
<tr>
<td>DEA</td>
<td>MACD (26,12,9)DEA</td>
<td>M1M2</td>
<td>M1M2</td>
</tr>
</tbody>
</table>

Optimal Parameters Selection

The impact of the ANFIS model on the input variables is important. In the number of input variables, too many input variables will lead to the calculation of the model increased time, at the same time, will increase the system instability. In the specific experiment process, if each input gives two membership functions, the total parameter quantity is 20, 44 or 96 for the 2, 3 or 4 inputs of the Gaussian membership function. The total parameter quantity in the triangle or bell membership function is 24, 50 or 104, given that we consider the case where the input variable is 3. And we choose all the possible digital combination, respectively, given the two Gaussian membership function after the iteration of the error. In the end, we selected the five combinations with the smallest error.
As shown in the figure, when the number of input variables is three, the error of model training is obviously decreased compared with the two input variables. Therefore, three input variables are most of the parameters are optimal when the iteration error is 30 times. The steps are not reduced; the iterative step size is set to 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, step size is canceled automatically and always fixed. Methods like Mean Square Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are used to test the accuracy of the prediction.

Experimental Results

Relationship between Filter-based Approaches

After selecting the best input variable, we classify the parameter optimization samples into two categories. One is the training sample; iteratively optimize the parameters according to the algorithm mentioned above. The second is to test the sample, not to optimize the parameters once, calculate the sample once under the error, and finally select the smallest error of this parameter as the final model. We selected the last 60 samples as a training sample, except for all the samples taken as training samples. According to the parameters determined above, a total of 21 training models and 5 input samples were generated. We always selected the five models with the smallest error as the input model and the average of 25 predicted values as the final prediction result.

In picture, we compared the prediction result with the actual value. The two curves represent the timing behavior prediction and the actual stock index change, respectively. The red curve represents the trade timing signal calculated from the results of the index forecast. When the forecast results of the day when the index is greater than 0, the issue of the long position signal, when the forecast day the results of the index is less than 0, issued short positions signal. The blue curve indicates the prediction of the HS300 index from the ANFIS with the stock index as the training target and the model based on the forecast data.
Figure 3. ANFIS marketing timing results.

We counted the correct number of times and the correct rate of the five optimal input combinations with the smallest error in the prediction results. It can be seen that the correct rate is maintained above 50% and we select the average of the results of these combinations as the result of the model output.

Table 2. Results of the model output.

<table>
<thead>
<tr>
<th>Input variable</th>
<th>Market Timing times</th>
<th>Wright times</th>
<th>Wright rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turn Volume\RSI\Close</td>
<td>276</td>
<td>163</td>
<td>59.071%</td>
</tr>
<tr>
<td>Turn Volume\RSI\MACD</td>
<td>276</td>
<td>159</td>
<td>57.609%</td>
</tr>
<tr>
<td>RSI\MACD\ROC</td>
<td>276</td>
<td>157</td>
<td>56.884%</td>
</tr>
<tr>
<td>RSI\MACD\Turn Volume</td>
<td>276</td>
<td>153</td>
<td>55.435%</td>
</tr>
<tr>
<td>Turn Volume\MACD\Low</td>
<td>276</td>
<td>148</td>
<td>53.623%</td>
</tr>
</tbody>
</table>

Analyzing the structural properties of the five different input samples from the results in the table, the accuracy of the forecast results is more than 50%, and from the predicted situation, the results of emotional values tend to predict the stock market is rising, the forecast is accurate. The rate is higher than the forecast rate of decline. This shows that the feelings of investors in the volatility of the stock market has a certain impact, and investors who are more emotional are relatively conservative, only when the stock market has a clear upward trend, the emotional value will tend to actively buy behavior.

The Empirical Results Comparison

In this part, we compared the predictions of ANFIS models and traditional time series models including Artificial Neural Networks and Support Vector Machine, and calculated their MSE, MAE, MAPE, respectively.

Figure 4. Error Comparison.
Through the comparison of the above results, we can think that ANFIS model has better prediction results, which not only combines the advantages of fuzzy system, but also has the characteristics of adaptiveness, can be very good to achieve the prediction of financial trends. Of course, the ANN and SVM models can also go for relatively good results, but overall, ANFIS has more accurate and stable predictive capabilities.

**Conclusions and Future Work**

The artificial intelligence method is applied to the field of stock price fluctuation. For example, artificial neural network, support vector machine model, and hidden Markov model, this paper uses the fuzzy reasoning mechanism to judge the price fluctuation method based on ANFIS model.

In this paper, an improved ANFIS model is proposed to construct a custom optimization target to set the opening threshold and stop line as the trading strategy, which makes the model improve the accuracy and simulation yield.

**References**


