Stock Market Trend Prediction with Sentiment Analysis based on LSTM Neural Network

Xu Jiawei, Tomohiro Murata

Abstract—This paper aims to analyze influencing factors of stock market trend prediction and propose an innovative neural network approach to achieve stock market trend prediction. With the breakthrough of deep learning recently, there occurred lots of useful techniques for stock trend prediction. This thesis aims to propose a method of feature selection for selecting useful stock indexes and proposes deep learning model to do sentiment analysis of financial news as another influencing factor influencing stock trend. Then it proposes accurate stock trend prediction method using LSTM (Long Short-term Memory).

Index: Stock trend prediction, LSTM, Sentiment Analysis, Deep learning, Chinese Stock market, Feature Selection...

I. INTRODUCTION

Stock market trend prediction plays a significant role in investment field. A lot of technique analysis methods occurred to solve this problem. With the stock market developing, traditional techniques can hardly achieve better performance now. With deep learning developing, a lot of techniques like LSTM, neural network proved to be effective in finance field. We aim to use deep learning method on stock trend prediction and analysis the influencing factors of stock trend prediction method based on LSTM neural network. This study mainly focused on feature selection, sentiment analysis of financial news and neural network structure.

II. BACKGROUND /RELATED LITERATURE / DATASETS

A. Background

Prediction of stock market has attracted attention from industry to academia. Various machine learning algorithms such as neural networks, genetic algorithms, support vector machine, and others are used to predict stock price. However, accuracy is unsatisfied, because of the reasons:

1. Data Noise:
There are lots of unprocessed factors. It causes some problems like data redundancy, data noise and overfitting.

2. Market Emotion:
Stock market is a stochastic field. Various aspects influence investors’ emotion. Market emotion strongly affects the stock market trend. And investors’ emotion is usually affected by financial news.

3. Time series Information:
Traditional methods can handle time series data, but with limited performance. And some methods like ARIMA have poor effect on big data.

B. Related Literature

Huyinh, Huy D., L. Minh Dang, and Duc Duong introduced a new prediction model depend on Bidirectional Gated Recurrent Unit (BGRU)[3]. Sun, Haonan[1] developed a predictive model to improve the accuracy by enhancing the denoising process which includes a training set selection based on four K-nearest neighbors (KNN) classifiers to generate a more representative training set and a denoising autoencoder-based deep architecture as kernel predictor. This paper has a shortcoming that he didn’t straightly use the time series data. Some researchers attempt using financial news data to solve some stock selection problem[4].

C. Datasets

This project mainly uses 2 different kinds of datasets: Financial news data for sentiment analysis, stock data. We got this data from web crawler and did lots of data preprocessing of them to be suitable format.

C.1 Sentiment analysis data

When an investor plan to do investment on some stock, he always read some financial news from website to get some advice from experts or reports. Proposed method is to do sentiment analysis of news to help investors modify their investment strategy. Following is financial news data preprocessing workflow:

![Financial news processing diagram](image)

Fig. 1. Financial news processing diagram.

After web crawler processing, financial news title data becomes like this:

等待非农就业数据 美股午盘维持小幅上扬
sentiment analysis of financial news and Deep Learning based trend prediction methods on stock data. In the feature engineering it has 2 steps: using auto encoder to do feature dimension reduction and comparing different feature combinations to find optimized features. In the sentiment analysis method of financial news, we labeled lots of financial news title text with pos/neg labels to represent their sentiment and then predict the newly published news data. In trend prediction model, I build LSTM type neural network for improving time series processing ability.

1. Feature Engineering

1) SDA (Stacked Denoising Auto Encoder) is applied to reduce the dimension of features which is not sensitive to the noise. [1] An autoencoder is a type of artificial neural network used to do unsupervised learning of data coding. The aim of an auto encoder is to learn higher-level representation for a set of data, typically for dimension reduction.

2) Then different sequence length and different sample amount are applied to experiment for finding an optimal sequence length [9].

2. Sentiment Analysis of Financial News

LSTM neural network is applied for sentiment analysis. Input is sentence vectors. Then split data into several parts by different companies. Put them into LSTM layer. Then add a dense layer. Output layer will output sentiment analysis result, value range from 0.0~1.0. And then a 4-quantile value is used. If value <25%quantile as -1(negative); 25%quantile =< value <= 75%quantile as 0(even); value > 75%quantile as 1(positive). Model design is showing below:

III. APPROACH

Main approach contains 3 parts: feature engineering, trend prediction, and sentiment analysis.
Sentiment analysis results + stock data are input of LSTM Model [2]. And output is the trend of stock price movement.

After feature selection and sentiment analysis, we got stock data at day t and sentiment label at day t. Combine them as the input. Output is trend prediction result -1/0/1 representing negative/even/positive. The framework shows below:

![LSTM neural network prediction model](image)

**IV. EXPERIMENTS**


**Experiment 1: Accuracy Evaluation**

**Objective:** Compare proposed method with other commonly used prediction methods using same experiment environment.

**Process:**
1. Using different prediction methods to predict stock trend and compare with proposed model. 30 days stock data as input and predict average stock trend of next 3 days.
2. Calculate average change price of next 3 days using change price:
   - Get 3 quantile value 0.33AC, 0.67AC, and convert average change price into category labels:
     1. If C < 0.33AC: label = -1 (Down)
     2. If 0.33AC < C < 0.67AC: label = 0 (Even)
     3. If C > 0.67AC: label = 1 (Up)
3. Training model, stock data as input matrix, real_trend as output and predict pred_trend.
4. Test data: I use stock data of Ping-an Finance Group (stock code: 601318.SH) from 2011/01/01~2018/11/18:

<table>
<thead>
<tr>
<th>date</th>
<th>open</th>
<th>high</th>
<th>low</th>
<th>close</th>
<th>pre_close</th>
<th>change</th>
<th>pct_chg</th>
<th>vol</th>
<th>amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011/01/04</td>
<td>56.85</td>
<td>57.50</td>
<td>56.60</td>
<td>56.91</td>
<td>56.18</td>
<td>0.75</td>
<td>1.34</td>
<td>248626.32</td>
<td>1400409.336</td>
</tr>
<tr>
<td>2011/01/05</td>
<td>54.54</td>
<td>54.94</td>
<td>54.54</td>
<td>54.94</td>
<td>54.84</td>
<td>-2.06</td>
<td>-3.46</td>
<td>42754.25</td>
<td>2366305.774</td>
</tr>
<tr>
<td>2011/01/09</td>
<td>54.94</td>
<td>54.94</td>
<td>54.54</td>
<td>54.54</td>
<td>54.84</td>
<td>-2.27</td>
<td>-4.14</td>
<td>547076.65</td>
<td>4971801.955</td>
</tr>
</tbody>
</table>

Fig. 4. Ping-an Finance Group stock data.

Then, I calculate average change price of next 3 days and named a new column ‘ave_change’ as prediction target. Experiment results shows below:

![Stock trend prediction of Proposed method and SVR](image)

And then, I get 3 quantile value 0.33AC, 0.67AC, and convert ave_change price into category labels:
1. If C < 0.33AC: label = -1 (Down)
2. If 0.33AC < C < 0.67AC: label = 0 (Even)
3. If C > 0.67AC: label = 1 (Up)

Then, I calculate average change price of next 3 days and named a new column ‘ave_change’ as prediction target. Experiment results shows below:

![Trend prediction Accuracy by Proposed Method](image)

![Trend prediction Accuracy by SVR](image)

Fig. 6. Trend prediction Accuracy by Proposed Method

Points in the graph means prediction is true or false.
(Point value = prediction trend result – real trend result)
Y axis means point value: 0 is true, 1 and -1 is slightly wrong, 2 and -2 is totally wrong.

Proposed method accuracy is 65.78% while SVR is 40.31%. Random selection accuracy is 33%. Both have better performance than random selection and proposed method is the best.
Experiment 2: Feature Selection Effect

Objective:
Compare different feature combination and find optimized feature selection. Furthermore, evaluate feature dimension reduction’s performance. Stock data is from China Vanke Co. Ltd. (code: 000002.SZ) from 2011/07/01~2018/11/18. And A-share composite stock data (code: 000001.SH). We establish the following 5 comparison models experiments to find the optimal combination.

Technique Index:
- Macro Index (Shanghai Composite Stock price):

Discussion:
1. M2 didn’t increase accuracy compared with M1, but it causes more feature which will cause more computing cost. Technique index has little effect on trend prediction. Technique index is redundant to fundamental index in this task.
2. M3 is better than M1. Macro index has big effect. We can say individual stock price has positive correlation with whole stock price.
3. From M4 we find accuracy decreased. With too much features cause gradient vanishing and data redundancy.
4. Compare M5 with M4, we find using Auto-encoder for dimension reduction can solve gradient vanishing and improve prediction performance.

Experiment 3: Sentiment Analysis Performance Evaluation

Objective:
Compare with and without Sentiment analysis. I have established text sentiment analysis tool which is effective on finding sentiment of financial news. And then I calculated sentiment index of each day by sentiment news sentiment statistics. Stock data is China Composite Stock index. Shanghai Composite Stock Index(000001.SH) from 20160401 ~ 20171001. News data is from a stock forum named GUBA news data related to Shanghai composite index from 20160401~20171001.

Then convert them into sentiment label for each day:

Table 4. Trend prediction accuracy after feature selection

<table>
<thead>
<tr>
<th>No.</th>
<th>Model</th>
<th>Dimension</th>
<th>Feature combination</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M1</td>
<td>9</td>
<td>1.Stock data</td>
<td>60.85%</td>
</tr>
<tr>
<td>2</td>
<td>M2</td>
<td>20</td>
<td>1.Stock data + 2. Technique index</td>
<td>60.55%</td>
</tr>
<tr>
<td>3</td>
<td>M3</td>
<td>18</td>
<td>1.Stock index + 3. Macro Index</td>
<td>66.36%</td>
</tr>
<tr>
<td>4</td>
<td>M4</td>
<td>29</td>
<td>1.Stock data + 2. Technique index + 3. Macro Index</td>
<td>61.46%</td>
</tr>
<tr>
<td>5</td>
<td>M5</td>
<td>18</td>
<td>1.Stock data + 2. Technique index + 3. Macro Index + 4.Auto-encoder</td>
<td>64.83%</td>
</tr>
</tbody>
</table>

Then convert them into sentiment label for each day:

Fig. 8. Stock trend prediction with Sentiment analysis and without

Fig. 9. Trend prediction accuracy with sentiment analysis

After sentiment analysis, each news data will have a result like below:

<table>
<thead>
<tr>
<th>date</th>
<th>trade_date</th>
<th>sentiment_label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 20160401</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>1 20160402</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>2 20160403</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3 20160404</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>4 20160405</td>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>

Final experiment results showing below:
Fig. 10. Trend prediction Accuracy without sentiment analysis
In first time experiment, prediction accuracy increased to 62.5% from 54.28% with sentiment analysis. I repeat the experiment for 6 times. Generally, Sentiment analysis has positive effect.

<table>
<thead>
<tr>
<th>No.</th>
<th>No Sentiment</th>
<th>With Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.28%</td>
<td>62.5%</td>
</tr>
<tr>
<td>2</td>
<td>56.13%</td>
<td>56.94%</td>
</tr>
<tr>
<td>3</td>
<td>52.99%</td>
<td>58.88%</td>
</tr>
<tr>
<td>4</td>
<td>53.66%</td>
<td>48.62%</td>
</tr>
<tr>
<td>5</td>
<td>48.28%</td>
<td>57.45%</td>
</tr>
<tr>
<td>6</td>
<td>54.25%</td>
<td>66.32%</td>
</tr>
</tbody>
</table>

Table 5. Sentiment analysis effect experiment results.

Experiment 4: Stability and Generality of the propose method

Objective: Evaluate stability of proposed method on different market industry fields
Method: Stock data of different individual companies are used to do experiments separately and check the variance of prediction to find whether it’s a stable method.
Results showing below:

<table>
<thead>
<tr>
<th>Stock Num</th>
<th>Acc</th>
<th>Stock Num</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>000002万科地产</td>
<td>62.96%</td>
<td>600000浦发银行</td>
<td>71.65%</td>
</tr>
<tr>
<td>600028中国石化</td>
<td>68.15%</td>
<td>601166兴业银行</td>
<td>68.60%</td>
</tr>
<tr>
<td>600036招商银行</td>
<td>63.75%</td>
<td>601901方正证券</td>
<td>76.82%</td>
</tr>
<tr>
<td>600519贵州茅台</td>
<td>53.42%</td>
<td>600104上汽集团</td>
<td>65.58%</td>
</tr>
<tr>
<td>601288农业银行</td>
<td>51.84%</td>
<td>601088中国神华</td>
<td>70.10%</td>
</tr>
<tr>
<td>601318中国平安</td>
<td>65.52%</td>
<td>601668中国建筑</td>
<td>57.48%</td>
</tr>
<tr>
<td>601398工商银行</td>
<td>62.36%</td>
<td>000133美的集团</td>
<td>65.00%</td>
</tr>
<tr>
<td>601628中国人寿</td>
<td>64.73%</td>
<td>601998中信银行</td>
<td>71.42%</td>
</tr>
<tr>
<td>601857中国石油</td>
<td>67.26%</td>
<td>601939建设银行</td>
<td>61.31%</td>
</tr>
<tr>
<td>601988中国银行</td>
<td>66.56%</td>
<td>601857中国石油</td>
<td>70.26%</td>
</tr>
</tbody>
</table>

Table 6. Trend prediction accuracy of different individual stocks.

V. CONCLUSION

From the research, we proposed an accurate stock trend prediction method and convinced market emotion is a very important factor influencing stock market and can help improve prediction accuracy.

Proposed novel method with feature compression and sentiment analysis for stock trend prediction and it improved accuracy than SVR by about 20%.

Market emotion was caught by sentiment analysis and it is a very important factor influencing stock market and improve prediction accuracy by 5%.

Recurrent Neuron Network with LSTM (Long Short-term Memory) can handle financial time series data better than traditional time series prediction method.

My method is stable in different tasks like individual stock or composite stock prediction.

REFERENCES