Predicting Stock Index Using Multi-Source Data

Yuhan Su, Ruichuang Cao, Wei Xu

Institute for Interdisciplinary Information Sciences

Introduction

- Predicting changes in the stock market is important both for research and practice. From the very beginning, efficient market hypothesis says that the stock price is unpredictable, but numerous studies have shown that it could be predicted to some degree. At the same time, behavioral economics tells us that the sentiments of the pubic may be correlated with the economic indicator.

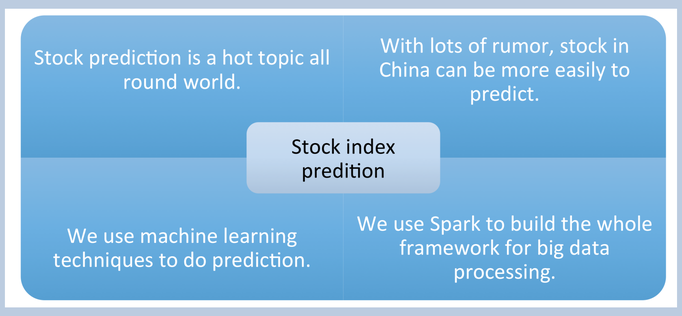

Figure 1: Project Introduction

- We are going to design and implement a stock price predicting system using social data and trading history data. We aim to get a higher prediction precision by using better algorithms and machine learning models, as well as speed up training and predicting methods by using Spark and HDFS. We plan to find the relationship between social data features and stock prices changes, train the predicting models and evaluate the model with traditional technical analysis methods.

Project Goals

The main goal of our project is to predict stock price or stock index variation tendency, which includes following sub-goals:

- Acquiring time series quantitative data and natural language data. Some of them could be download from public datasets, and some could be crawled from website.
- Generating training features from natural language data source and history trading data. The latter one is simply by automatically selecting significant factors according to experiments' feedback. But the former one meed generating specialized sentiment dictionary for different data source.
- Using time series features to train different models that best describe the relationship between prediction target and features.
- Combine the results of different models to give the final prediction result. We will try different combining methods such as AdaBoost.

Data Source

The data source we now have includes:

- User generated information. Including posts and comments from Sina Guba, browsing history from Weixin. Figure 2 is a schematic sentiment result, which shows the numerical results that we can generate from these data.

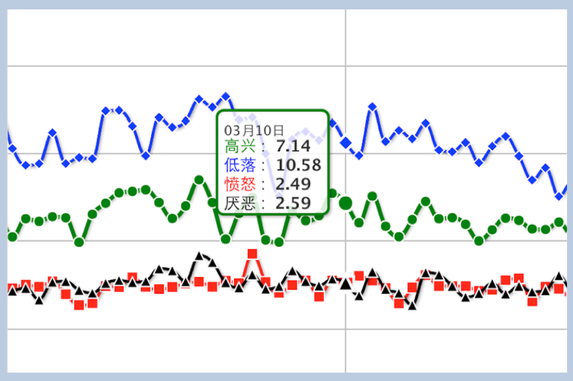

Figure 2: Sentiment Sample*

- Daily trading history for SSE50 and its stocks. We choose SSE50 index as our predicting target. We get these data using the python lib named Tushare and download some from Sina Finance.
- High frequency trading history for some stocks. In one second it may contains several items, in which there are data for many kind of prices and amounts, including buying/selling price/amount order.**

Data Pre-Processing

- For natural language data, data pre-processing includes manually labeling the sentiment for key word, generating the sentiment dictionary for different data source, and transforming them into numerical data. We may generate features as Fig 3 shows.

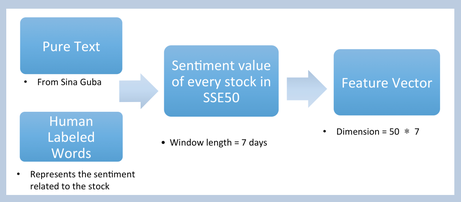

Figure 3: Natural Language Data

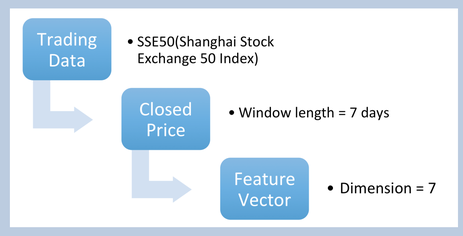

Figure 4: Trading History Data

- For trading history data, Fig 4 shows a basic generating procedure. Then we can choose the stock price of the next day as predicting target.

System Architecture and Processing

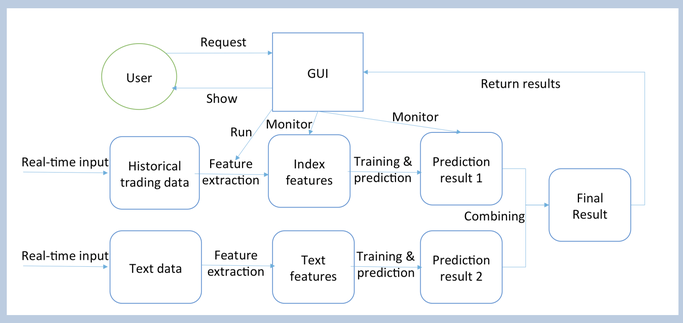

Figure 5: System Architecture

- When the user requests for prediction, the system will parse features from various data source stream and give the final prediction result. The training model will also be updated with the new data. Users can monitor the whole process on a web UI. They can change the model parameters, modify the codes, and monitor the data stream from the input to the final result.

Prediction Results

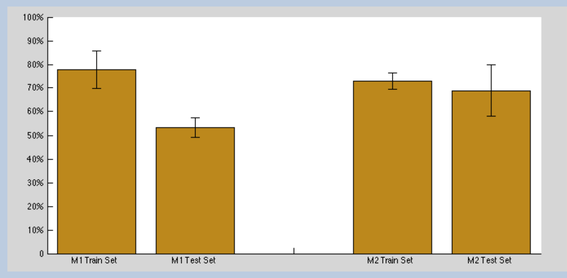

Figure 6: Experiment Result

- We formulate the problem as a regression problem. Then we calculate the accuracy by comparing whether the stock index rises or falls, like a binary classification. Here is the precision of a sample result for SSE50 using linear regressing and random forest. In future we will try other supervised algorithms to improve the result.

Conclusion

- From the preliminary result, we can find that the stock index is predictable to some degree.
- In future, we will try to improve feature extraction, model selection and combine other data source. We will also implement the web UI design.

Acknowledgments

- Thank IBM China Research Lab for providing computing resources on SuperVessel and give important instructions on weekly meeting.
- *Emotion Sample Figure from XinQingSouSuo
- **Thank Baitao Long for sharing trading data and give useful suggestions.

Contact Information

- Web: http://iiis.tsinghua.edu.cn
- Email: Yuhan Su syhmartin@yeah.net
  Ruichuang Cao create0818@163.com
  Wei Xu wei.xu.0@gmail.com