Machine Learning Final Project Proposal

Liang Shuailong 1000829

Task: Stock Market Prediction

1. Background

It is possible to predict stock price based on news data. Previous method to predict stock market uses shallow features such as Bag of Words, named entities and noun phrases. However, these features fail to capture the structure of event which is important to stock market prediction. In Ding et al., 2014, he compared the performance of the combination of bag of words features, structured event features and SVM, deep neural network models, and get the performance with is better than the State-Of-Art system.

2. Papers

1) Using Structured Events to Predict Stock Price Movement: An Empirical Investigation.
2) Deep Learning for Event-Driven Stock Prediction

3. Tasks

The main task is to replicate Ding's experiments using BoW/Event features, SVM/Deep Neural Network models. The financial news datasets are from Reuters (106,521 docs) and Bloomberg (447,145 docs). The stock price data is S&P 500 index. Specifically, the tasks are arranged according to the steps below.

1) Data acquisition and feature extraction. In this stage, data will be collected from Reuters and Bloomberg. BoW features and Event feature will be extracted from the data. The classification of news to companies is also done in this stage.
   [March 4 - March 11] 7days

2) Use SVM to predict prices of several individual stocks using data and features extracted from step 1).
   [March 12 - March 19] 7 days

3) Build a Deep Neural Network to predict prices of several individual stocks.
   [March 20 – April 2] 14 days

4) Try different time spans: 1 day, 1 week and 1 month and compare the result.
   [April 3 – April 10] 7 days

5) Use Neural Tensor Network to train the event embedding to represent structured events to address the problem of sparsity.

[April 11 – April 18] 7 days

6) Write report and refine the experiments
[April 19 – April 22] 3 days

4. Risks

The risks may include that the expected accuracy may not be reached, the deep neural network may be difficult to train, the experiments may take very long time to run, etc.