Hybrid Attention Networks for predicting Chinese stock trend by news data

Anonymous Author(s)
Affiliation
Address
email

(Implementation Track, group of 4)

Abstract

Making a prediction of the trend of the stock market is a very important part for investors to maximize profit. However, predicting the trend of the stock market is a very difficult due to the sensitivity of the stock market. The stock market is very sensitive to the real-world, there are many factors that can influence the stock market, for example, the sudden break of the pandemic, the bankruptcy of a relative company, etc. Internet, as a huge source of information, can be an exhaustive description of what is happening in the world. However, it is important to distinguish the high quality and real knowledge in the lake of information. In order to address these challenges, we implemented and improved methods from Peking University and Microsoft research team which used news data to predict the stock trends and improve the annualized return. They purposed a framework with a Hybrid Attention Networks (HAN) and a self-paced learning mechanism to address three main problems which are *sequential content dependency, diverse influence, and effective and efficient learning* when dealing with online stock news. In this paper, we are going to replace the bi-directional GRU-based RNN method with BERT and train with both stock news data and previous stock price data, which can ensure a better prediction performance.

1 Introduction or Problem statement

Predicting the future trend of the stock price is always a goal that all investors doing in order to make the profit maximized. However, this is a very challenging task due to the high sensitivity of the market which made it very fluctuated. There are conventional ways that trying to use information from price, volume, and other fields to predict the future price, but these methods cannot reveal the ground truth beneath the fluctuation of the stock price, which made them limited. So we come up with an idea that, if we can gather all information about certain companies, we can make a strong prediction about their stock prices. The Internet, as a huge pool of information, can be considered to hold all recent information in the world. In other words, if we want some public accessible information, it must be able to be found somewhere on the internet. This job requires people to search and extract useful and reliable information from the internet. Fortunately, as the Natural Language Processing techniques and Web crawling techniques improved, which made our goal become realizable. The research team stated that there are several ML tools based on analyzing the Wall Street journal. So they decided to develop a tool called Hybrid Attention Networks (HAN) to predict stock price by relevant news.[1] Our project is focusing on implement HAN in the Chinese stock market and evaluates its performance. Since BERT is one of the state-of-art nlp techniques being widely used today, we also want to replace the bi-directional GRU structure in the HAN framework with BERT to enhance the model prediction performance.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2 Method

The core idea is to imitate the learning process of human when facing chaotic online news which can be concluded into three characteristics:

- Sequential Context Dependency: It refers to the fact that a single news is more informative within a broader context than isolated. Therefore, when we take a sequence of related news as a unified content, we can make better prediction.
- Diverse Influence: Different online news have different influence. It is common that one critical news can affect the stock price for weeks, whereas a trivial one may have zero effect. So when we predict the subsequent stock trend, we need to comprehensively identify the estimated impact of each online news.
- Effective and efficient Learning: It is not surprising that the content of news contains vague information on stock trend which leads to the fact that we can not make any reliable new-oriented prediction. In order to address this problem, people usually tend to first gain knowledge from the more common situations before turning to more complicated situations. And if there is limited number of stock news in a period, we may discard those data due to the efficiency requirement.

Ziniu Hu et al. proposed a Hybrid Attention Networks (HAN) [1] to capture first two characteristics, Sequential Context Dependency and Diverse Influence. The Hybrid Attention Networks (HAN) includes temporal-level attention-based recurrent neural networks (RNN) to identify more influential time periods of the sequence and news-level attention-based recurrent neural networks (RNN) to identify more influential news at the same time point. With structure of recurrent neural networks (RNN), we are able to consider a sequence of related recent news as a unified context and the attention mechanism enables the process of identifying diverse influence of different news and different time periods.

To perform a more effective and efficient learning, they implemented a Self-Paced Learning (SPL) algorithm. Kuma et al. designed Self-Paced Learning (SPL) which is based on Curriculum learning but it can optimize the original objective and curriculum design at the same time. It skips certain data points that are considered to be yet too hard and gradually increase the complexity of training samples thus the model performs better predictions.

Specifically, the framework of Hybrid Attention Networks (HAN) includes:

1. a pre-trained unsupervised Word2Vec as the word embedding layer;
2. an attention layer to catch the unevenly distributed importance level of a sequence of stock news in a day;
3. a bi-directional GRU network to incorporate information of a sequence of stock news to get a latent vector for one day;
4. Another attention layer to adjust the importance of each day's latent vector;
5. A standard Multi-layer Perceptron (MLP) to output the prediction results;

3 Related work

In the traditional finance theory, Brownian motion and Efficient Market Hypothesis (EMH) [2] are the cornerstones of the stock market prediction and strategies. However, many empirical results indicate that EMH is not applicable to the stock market. To capture the characteristics of time-series historic market data, neural networks have been considered to be a very effective learning algorithm for decoding those nonlinear time series data [3]. Among different widely used deep learning neural networks architectures, Recurrent Neural Networks (RNN) are best fitted for time series modeling problems in financial market. In 1996, Roman et al. [4] used back propagation and RNN models to predict stock index among different stock market. Recently, Long-Short Term Memory (LSTM) is considered as one of the most popular approaches for stock market prediction since it has the capability of holding past information [5].

Ziniu Hu et al. stated that although there has been a recent surge of interest on development of Deep Neural Network (DNN) architectures based on news or public data to predict stock trend movements [6], most of them are devoted to reliability and quality of information sources. In addition, Ronaghi et al. [7] designed s a Noisy Deep Stock Movement Prediction Fusion framework (ND-SMPF) for stock price movement prediction to address this problem.

4 Experiments

4.1 Implementation

To collect new data for our project, we use an amazing stock data crawling tool called Tushare. For model implementation part, in general, we are going to implement the Hybrid Attention Networks (HAN) framework mainly with PyTorch module. And regarding the self-paced learning (SPL) algorithm, we might need to write our own code to implement the parameter optimization steps in Python.

More specifically, for the deep learning network construction part, we need to define all the layer listed above and try running it on Google Colab which offers us a free GPU. For the self-paced learning part(SPL), we need to implment the iterative algorithm for parameters optimization such as back-propagation on our own. So we have to be careful about deriving the math formula correctly before coding.

4.2 Dataset

The data we are going to use for this project can be mainly divided into two parts: one is the relevant stock market news information, the other is the previous stock market prices. We retrieve our data via the crawling tool called Tushare, basically, it includes:

1. Historical stock open price, close price, all interesting factors for each stock
2. Corresponding news for each stock with published date
3. all the quick news from 2019-01-01 till now, still need to assign to corresponding stocks with names or ids
4. Overall news about finance or stock market with published date
5. all the general news from 2019-01-01 till now

This process is challenging because first we need to extract the publication timestamp, title, and content for each news. Then we need to correlate each of the collected news to a specific stock if the news mentioned the name of the stock in the title or content. We then filter out the news without any correlation to stocks. After such a process, for each stock, we then aggregate all the news in a certain date to construct the daily news corpus.

4.3 Evaluation

To specify the label of the tri-classification problem (UP, DOWN, STAY), we set up two particular thresholds to bin the rise percents (rise percent is defined as the increasing rate between next day's and today's open price), We define the thresholds so that the three categories are approximately even. We also split the dataset into a training set (66.7%), and a test set (33.3%). Then we further randomly sample a validation set from the training set with 10% size of it, in order to optimize the hyper-parameters and choose the best epoch.

In the experimental setting of a tri-label classification problem, as the three label is approximately evenly split, we choose accuracy, which means the proportion of true results among the total number of testing samples, as the evaluation metric.

5 Plan of Project

5.1 Anticipated division of work over team

For the model construction part: we are going to split our work mainly into 4 parts:

1. data crawling
2. HAN framework implementation(replace bi-directional GRU with BERT)
3. SPL iteration math formula derivation
4. SPL algorithm implementation in Python

For the actual experiment part: we are going to try 4 other models or deep learning frameworks to compare the performance with the improved HAN+SPL model described above. Each member would take care of one model or framework.

5.2 Experience or expertise that we might have to complete the project successfully

Two of our members have taken the NLP course which could help implement the BERT model here. And all of us have prior experience coding with PyTorch module. Besides, one of our members majoring financial engineering and risk management degree also has lots of experience working with stock market data.

5.3 Expected challenges and difficulties

The first challenge is to retrieve the data from the internet by using crawling tool. We have to learn how to use the new tool Tushare in a short time. And then we need to process the data by relevant each stock market news to a certain date, which is also a challenging work. For the model constrcution part, we need to not only implement the HAN framework but also replace the bi-directional GRU with BERT. Last but not least, we need to write our own code to implement the SPL algorithm.

5.4 milestones and contingency plans to get positive results

We are going to hold a weekly recurrent meeting to mainly focus on each part of work described above.

1. data crawling and pre-processing
2. HAN framework implementation(replace bi-directional GRU with BERT)
3. SPL iteration math formula derivation
4. SPL algorithm implementation in Python
5. experiment with new data and compare to other models' performance

References

[1] Hu, Ziniu, et al. "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction." Proceedings of the eleventh ACM international conference on web search and data mining. 2018.

[2] Malkiel, Burton G. "The efficient market hypothesis and its critics." Journal of economic perspectives 17.1 (2003): 59-82.

[3] Batres-Estrada, Bilberto. "Deep learning for multivariate financial time series." (2015).

[4] Roman, Jovina, and Akhtar Jameel. "Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns." Proceedings of HICSS-29: 29th Hawaii International Conference on System Sciences. Vol. 2. IEEE, 1996.

[5] Jia, Hengjian. "Investigation into the effectiveness of long short term memory networks for stock price prediction." arXiv preprint arXiv:1603.07893 (2016).

[6] Ding, Xiao, et al. "Deep learning for event-driven stock prediction." Twenty-fourth international joint conference on artificial intelligence. 2015.

[7] Ronaghi, Farnoush, et al. "ND-SMPF: A Noisy Deep Neural Network Fusion Framework for Stock Price Movement Prediction." 2020 IEEE 23rd International Conference on Information Fusion (FUSION). IEEE, 2020.