A LSTM-based method for stock returns prediction : A case study of China stock market

Kai Chen
Shanghai Jiaotong University
Shanghai, China
kchen@sjtu.edu.cn

Yi Zhou
Shanghai Jiaotong University
Shanghai, China
zy_21th@sjtu.edu.cn

Fangyan Dai
MD Anderson Cancer Center
Houston, USA
fdai@mdanderson.org

I. INTRODUCTION

Prediction of stock market has attracted attention from industry to academia [1, 2]. Various machine learning algorithms such as neural networks, genetic algorithms, support vector machine, and others are used to predict stock prices.

Recurrent neural networks (RNNs) are a powerful model for processing sequential data such as sound, time series data or written natural language [3]. Some designs of RNNs were used to predict stock market [4, 5].

Long Short-Term Memory (LSTM) [6] is one of the most successful RNNs architectures. LSTM introduces the memory cell, a unit of computation that replaces traditional artificial neurons in the hidden layer of a network. With these memory cells, networks are able to effectively associate memories and input remote in time, hence suit to grasp the structure of data dynamically over time with high prediction capacity.

The presented paper modeled and predicted China stock returns using LSTM. The historical data of China stock market were transformed into 30-days-long sequences with 10 learning features and 3-day earning rate labeling. The model was fitted by training on 900000 sequences and tested using the other 311361 sequences. Compared with random prediction method, our LSTM model improved the accuracy of stock returns prediction from 14.3% to 27.2%. Our efforts demonstrated the power of LSTM in stock market prediction in China, which is mechanical yet much more unpredictable.

II. OUR METHOD

In our LSTM model for stock prediction, one sequence was defined as a sequential collection of the daily dataset of any single stock in a fixed time period (N days). The daily dataset describes the performance of the stock with sequence learning features like closing price, trade volume on one particular day in these N days.

We labelled the performance of the sequences based on the earning rate, which was calculated by comparing the average closing prices in 3 days after the sequence with that of the last day in the current sequence.

Our model is composed of (1) a single input layer with the number of memory cells as that of the sequence learning features one sequence may hold, followed by (2) multiple LSTM layers and (3) a dense layer and (4) a single output layer with the number of memory cells as that of the categories of the sequence performance, which was seven as practically determined in this study.

The code of Keras[1] as following:

```
model.add(LSTM(max_feature, nu,
return_sequences=True))

model.add(LSTM(nu, nu, return_sequences=True))

model.add(LSTM(nu, nu))

model.add(Dense(nu, nb_class))

model.add(Activation('softmax'))

model.compile(loss='categorical_crossentropy',
optimizer=RMSprop)
```

It is essentially critical to choose sequence learning features in LSTM. There are at least four types of stock data: (1) the historic price data of the stock (e.g. volume, high, low, open); (2) the technical analysis data that is calculated from (1) (e.g. moving average convergence / divergence (MACD)); (3) the historic price data of market indexes and/or other related stocks; (4) the economic fundamentals (e.g. gross domestic product (GDP), oil price). Our sequence learning features used the 1st and 3rd but not 2nd type of stock data, in an effort to avoid the co-founding pitfalls, with also a limit of not including the 4th type of data. We expect the sequence learning power of LSTM would find us the best parameters. In current study, one sequence has 30 days of stock data and each daily data has 10 features.

III. EXPRIMENTS

- **Raw Data**: We acquired the historical stock data for China stock market in Shanghai[2] and Shenzhen[3] from Yahoo finance[4] (daily record of high / low / open / close / volume). It has 7767102 daily records of 3049 stocks from 1990/12/19 to 2015/09/10.

[1] http://keras.io/

[2] http://www.sse.com.cn/assortment/stock/home/

[3] http://www.szse.cn/main/marketdata/jypz/colist/

[4] http://table.finance.yahoo.com/table.csv?s=000001.ss

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

- **Sequence data**: One sequence contained 30 consecutive daily stock data. We got 1211361 sequences from 2013/06/01 to 2015/05/31. We used 900000 sequences (from 2013/07/01 to 2014/11/12) for training purpose and 311361 sequences (from 2014/11/12 to 2015/05/31) for validation.

- **Categorization of stock return**: Based on the earning rate, the sequence fallen into seven categories as defined by the following range: [,-1.5], [-1.5,-0.5], [-0.5, 0.4], [0.4, 1.4], [1.4, 2.5], [2.5, 4.3], [4.3,]. Here, for example the [0.4, 1.4] category contained the sequences whose earnings rate was between 0.4% and 1.4%. These ranges also ensured comparable number of training sequences in each categories

- **Training Detail**: We trained model by the stochastic gradient descent using RMSprop, with learning rate of 0.001. We used minibatch size of 64, and normalization was applied for each vector of a sequence by using the mean and standard deviation of each stock computed from the training set. We used a PC server as training platform (CPU: E5-2620, 64G memory, GPU Titan), we chose CentOS 7, Theano[5] and Keras as deep learning platform. The duration of one epoch is about 1500 seconds,

- **Compared Methods: Random**: Prediction by chance; **M1**: LSTM with closing price and trade volume as learning feature; **M2**: M1 with normalization; **M3**: M2 with extra 3 features added (High, Low, Open); **M4**: M3 with extra 5 features added (High, Low, Open, Close, Volume of Shanghai Securities Composite Index (SSE Index (000001.SS)); **M5**: only stocks of SSE ETF180 Index used for training and validation. We summarized results of different methods in **Table 1**.

Table 1. Results of Stock Return Prediction

Methods	Features	Accuracy
Random	NA	14.3%
M1	Close, Volume	15.6%
M2	Close, Volume	19.2%
M3	High, Low, Open, Close, Volume	20.1%
M4	SSE Index (Close, High, Low, Open, Volume), High, Low, Open, Close, Volume	24.1%
M5	SSE Index (Close, High, Low, Open, Volume), High, Low, Open, Close, Volume	27.2%

Our results above revealed that normalization was very useful for improving accuracy (19.2% vs 15.6%). The SSE Index further increased the accuracy (24.1% vs 20.1%), consistent with the notion that the market indexes affects stock return. We got better accuracy if using only stocks picked by Shanghai Securities ETF180. This suggests different stock sets will affect the accuracy of the prediction and it is necessary to separately run the prediction for different type of the stocks.

Albeit not a very satisfying accuracy, we found we could still improve it especially if we set the right threshold to effectively exclude the sequences with extremely low or high earnings rate. This turned up very helpful when choosing the stocks for analysis.

Our initial efforts demonstrated the power of LSTM in sequence learning for stock market prediction in China, which is mechanical yet much more unpredictable. This inspires us a lot more continuous and interesting works. For example, to include the MACD and other features in the learning feature sets and evaluate their contribution, analyze the stocks type by type, time window by time window due to the volatilities of the market indexes. We also plan to test more data as learning features including the international market indexes, price of bulk commodity as well as the breaking financial news and even the mood of the social network [7, 8].

ACKNOWLEDGMENT

The work is partially supported by the National Natural Science Foundation of China (Grant No. 61221001), and Shanghai Science and Technology Committees of Scientic Research Project (Grant No. 14DZ1101200).

REFERENCES

[1] Krollner, Bjoern, Bruce Vanstone, and Gavin Finnie. "Financial time series forecasting with machine learning techniques: A survey." (2010).

[2] Agrawal, J. G., V. S. Chourasia, and A. K. Mittra. "State-of-the-art in stock prediction techniques." International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2.4 (2013): 1360-1366.

[3] Lipton, Zachary C. "A Critical Review of Recurrent Neural Networks for Sequence Learning." arXiv preprint arXiv:1506.00019 (2015).

[4] Saad, Emad W., Danil V. Prokhorov, and Donald C. Wunsch. "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks." Neural Networks, IEEE Transactions on 9.6 (1998): 1456-1470.

[5] Rather, Akhter Mohiuddin, Arun Agarwal, and V. N. Sastry. "Recurrent neural network and a hybrid model for prediction of stock returns." Expert Systems with Applications 42.6 (2015): 3234-3241.

[6] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[7] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.

[8] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) 27.2 (2009): 12.

[5] http://deeplearning.net/software/theano/