CS501 Research Report

(reproduce machine learning approaches applied in predicting the stock price)

Due to the chaos and high volatility of stock behavior, the investment risk in the stock market is high. In order to minimize the risk, advanced knowledge of future stock price movements is needed. Traders are more likely to buy stocks whose value is expected to increase in the future. On the other hand, traders may not buy stocks that are expected to fall in value in the future [1]. Therefore, there is a need to accurately predict trends in stock market prices in order to maximize capital gains and minimize losses.

Luckyson etc. [2] applied random forest to generate a model to predict the direction of stock market prices. They proposed a new method to minimize the risk of stock market investment by using a powerful machine learning algorithms known as ensemble learning. The learning model is an ensemble of multiple decision trees and the predictive output of the model can be used to support the decisions of those who invest in the stock market.

However, there are some drawbacks in the random forest: the heuristic learning rule does not effectively minimize the global training loss; the model size is usually too large for many real applications [3, 4]. Ren etc. [5] proposed a global refinement approach to address these issues. The global refinement approach relearns the leaf nodes of all trees under a global objective function in order to use the complementary information between multiple trees. The objective function of the global refinement random forest can be expressed as follows:

$$\min_{W} \frac{1}{2} \|W\|_F^2 + \frac{C}{N} \sum_{i=1}^{N} l(y_i, \hat{y_i})$$

$$s.t.\ \ y_i = W\Phi(x_i), \forall i \in [1, N]$$

$$\Phi(x) = [\phi_1(x), \dots, \phi_N(x)]$$

$$W = [w_1, \dots, w_N]$$

where $\phi_i(x)$ is the indicator vector of a tree, w_i is the leaf matrix, y_i is the prediction of the individual tree, and $\hat{y_i}$ is the ground truth.

I reproduced the algorithm in the paper [] with the traditional random forest and the global refinement random forest. Both models are tested on the historical stock prices of last five years, including Google, Apple, 3M and GM. From the results, we can find that the global refinement random forest model can achieve better performance almost in all metrics (precision, recall, f1, and accuracy). The results are given in the table below.

Google	Random Forest	Global Refinement Random Forest
precision	0.78	0.79
recall	0.92	0.93
f1	0.83	0.85
accuracy	0.81	0.83

Apple	Random Forest	Global Refinement Random Forest
precision	0.86	0.87
recall	0.91	0.91
f1	0.89	0.89
accuracy	0.86	0.86

3M	Random Forest	Global Refinement Random Forest
precision	0.78	0.79
recall	0.87	0.89
f1	0.82	0.84
accuracy	0.81	0.82

GM	Random Forest	Global Refinement Random Forest
precision	0.75	0.76
recall	0.88	0.92
f1	0.81	0.83
accuracy	0.8	0.82

Reference:
[1] L. Khaidem, S. Saha, S. R. Dey, Applied Mathematical Finance (2016).
[2] S. Boonpeng, P. Jeatrakul, 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI) (2016), pp. 1–6.
[3] T. Bylander, Machine Learning, University of California Berkeley, 48, 287 (2002).
[4] L. Breiman, Machine Learning, University of California Berkeley, 45, 5 (2001).
[5] S. Ren, X. Cao, Y. Wei, J. Sun, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015).