深度學習之應用於高頻交易中間價走勢預測

Deep Learning for Predicting Mid-Price Movement in High Frequency Trading

組別: B253

指導教授: 翁詠祿

組員姓名: 陳昭維 林禹陞

報告摘要

隨著資訊科學以及資料的公開和透明化，以及同時間硬體的進步以及運算能力的提升。推論統計學(Statistical Inference) 應用於解決現實生活問題以及專業領域的研究的實現也逐年提升。

其中，金融交易為大量使用推論統計學的產業和領域，隨著交易規模和速度的提升，傳統使用人為因素去判斷是否買賣金融商品的策略已逐漸被淘汰。而使用收集交易時的資訊並將其量化與特徵選取後制定演算法與選定買賣的標籤去實現自動交易已成為金融單位的主要策略。

本篇專題將以參考論文“ DeepLOB: Deep convolutional neural networks for limit order books”[1]作為實作的原理參考和模型建置基礎，對開源數據 FI-2010 dataset[2]的限價委託簿(Limit Order Book)數據進行前處理，利用委買賣資料獲得中間價以預測股市的走向而產生決策。其中的原理在於能夠利用一定數量的時刻點資訊預先獲得買賣時刻點所該進行的決策，領先其餘投資人進行交易，在金融市場中獲利。

透過參考論文之實驗理論之啟蒙以及對於金融時序列資料之專業知識，我們奠基於之上建立了深度學習模型來做為預測股市的架構。我們透過許多數學計算和深度學習網路的嘗試和訓練，驗證了方法的可行性，最終搭建出屬於我們的模型並且超越原作者[1]架構的預測能力以及延遲表現。

報告內容

1. 研究動機

在大學的課程中，我們學習基礎的知識和題材，但是我們往往只能從最表層的角度去了解應用面。作為電機系的學生，將來無論是朝向學界抑或是業界發展，我們會遇到的問題層面不會單純屬於電路設計或是軟體撰寫層面，我們會結合自身的專業於解決現實世界的問題。我們希望結合電機系的數學訓練在參考論文[1]的基礎之下，發揮想法以及實作精神，嘗試超越作者以及其他 state-of-the-art 的模型，並且學習利用科學思維解決人文問題。

2. 研究方法

研究相關理論，開源的證券限價簿成為至關重要的存在本專題以及參考論文[1]中共同對於 FI-2010 dataset[2]進行模型建置的輸入資料。此資料集為首個於高頻交易中公開的限價簿時序列資料，其中數據由納斯達克北歐交易所五間公司[2]的限價簿組成(表 1)。資料集的收錄時間為 2010 年 6 月 1 日至 2010 年 6 月 14 日。每日(平常日)的收錄時間為 10:30 至 18:00。

數據通過 ITCH flow 將資訊轉變成實作數據。ITCH flow 經由 C++ 將相關的文字、事件和交易行為都轉換為以日為單位經過分類的 h5 檔案。再來再經由 MATLAB 程式重新重組句子和限量委託簿以及其他的資訊。最終再經過機器學習得出最終實作使用資料集。

在獲得了特徵標準化的數據後，我們將數據中的委買/賣的價格和數量作為訓練的特徵，可以將 input 看為是二維的矩陣，而每個行則是由 $x_t = [p_a^{(i)}, v_a^{(i)}, p_b^{(i)}, v_b^{(i)}]_{i=1}^{n=10}$的向量為結構，其中 p 為價格(price)、v 為交易量(volume)、a 為賣出(ask)、b 為買入(bid)。我們使用 10 個時刻點的同結構向量作為列，因此輸入資料的維度如下所示:

$$\mathrm{X} = [x_1, x_2, \dots, x_{10}]^T \in \mathbb{R}^{10\times 40}$$

在資料標籤中，我們利用中價去標示價格漲跌的方向。中價定義為下:

$$p_t = \frac{p_a^{(1)} + p_b^{(1)}}{2}$$

若單純比較 p_t 以及 p_{t+k}的差異來判斷漲跌會失去資訊其中真正的意義。因此我們定義 m_-為前 k 個中價的平均值而定義m_+為後 k 個中價的平均值，公式如下:

$$m_- = \frac{1}{k}\sum_{i=0}^{k} p_{t-i} \qquad m_+ = \frac{1}{k}\sum_{i=0}^{k} p_{t+i}$$

我們又定義以下的公式:

$$l_t = \frac{m_+ - p_t}{p_t}$$

作為 t 時刻時的漲跌方向，我們又訂一個閥值α做為判別漲跌的參數:

$$price\ movement = \begin{cases} +1 \text{ ，} l_t > \alpha \\ -1 \text{ ，} l_t < \alpha \\ 0 \text{ ，} else \end{cases}$$

在數據測驗中我們使用了 5 種 k 預測，其中為 k = 10、20、30、50、100。

3. 模型總覽

我們所搭建的模型總參數量為 31,209 個，模型結構如**圖 1** 所示:

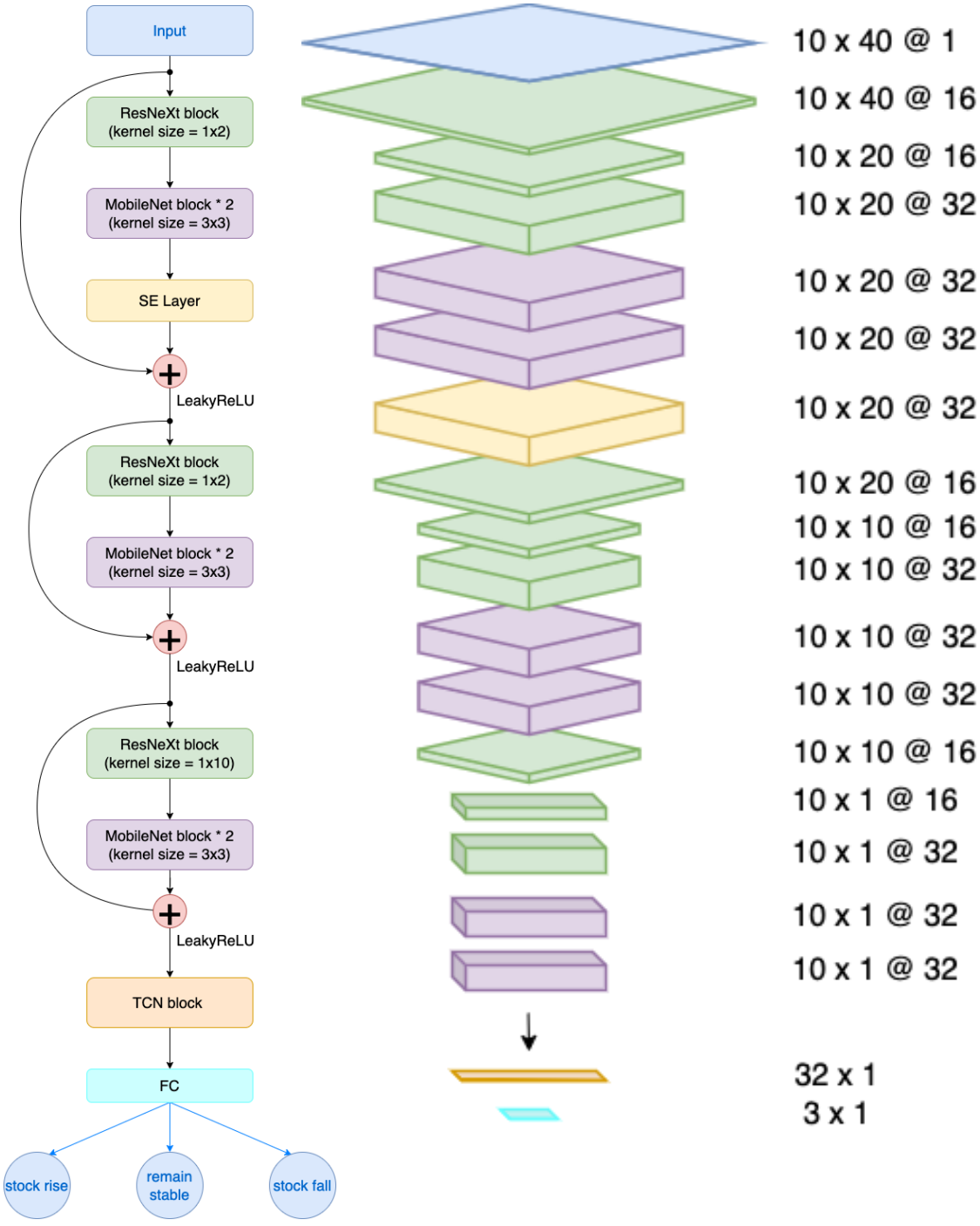

圖 1 本專題實作模型總覽

4. 研究結果

我們在實作當中與原作者使用相同的兩種測試方式作為 baseline 之間的比較。第一種測試方式我們使用前 x 天的資料訓練模型後使用第 x+1 天的資料於驗證模型，x 為 1 至 9 天。而第二種測試方式則為使用前 7 天的資料訓練後，去驗證後 3 天的資料。

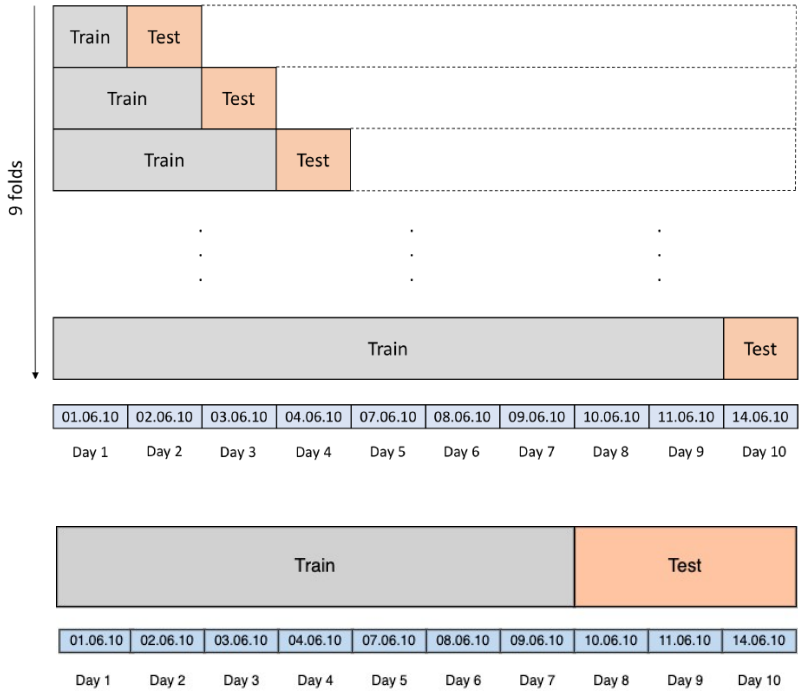

圖 2 兩種測試方式[1]

我們採取與參考論文[1]相同的驗證方式，採用比較 state-of-the-art 的方式來比較各模型對於 FI-2010 dataset 的四項指標:Accuracy、Precision、Recall 以及 F1-score。在原論文中與之比較的模型為 RR[2]、SLFN[2]、LDA[9]、MDA[9]、MTR[9]、WMTR[9]、MCSDA[10]、BoF[11]、N-BoF[11]以及 B(TABL)[12]和 C(TABL)[12]、SVM[13]、MLP[13]、LSTM[13]、CNN-I[14]、CNN-II[14]。

在第一種測試方式中，我們使用先前提到的 k 個中價來做為漲跌判斷的標籤，我們使用了 k = 10、50、100。而於第二種測試方式中我們使用了 k=10、20、50。

圖 3 和**圖 4** 中為我們的實作結果，其中表中黑字為原論文[1]提出與其模型比較之 state-of-the-art 模型表現，而橘字即為原論文[1]中所提出之模型，藍字為我們實作的模型。

從表格中可以看出我們無論是方法一或方法二中，使用的各項 k 值皆可以超越參考論文中的四項指標，這驗證了我們在於模型建置上的選擇是優於原作者的選擇。

我們能更加正確地捕捉了時序列之中的訊息，準確地預測出了股市漲跌之預測。

而除了精準度之外，我們在模型的參數量上也達到了 50%的減少，可以減少硬體的運算時間以及數據推論得出結果的時間延遲。

Latency 的計算方式考慮到實際情況會以 batch size 等於 1 的情況下送進 model 處理，因此雖然在訓練模型時我們是用 batch size 等於 32 來做訓練，但在測試 latency 時我們使用 batch size 等於 1 來量測，較符合真實情況。

圖 5 中為我們與其於 state-of-the-art 的參數之間比較，在參考論文[1]中，作者使用的硬體和我們測試時的硬體環境並非相同，因此我們還原作者的架構在相同的硬體下進行實作而得出**圖 5** 之數據表現。

Model	Acurracy %	Precision %	Recall %	F1 %
Prediction Horizon k = 10				
RR [2]	48.00	41.80	43.50	41.00
SLFN [2]	64.30	51.20	36.60	32.70
LDA [9]	63.83	37.93	45.80	36.28
MDA [9]	71.92	44.21	60.07	46.06
MCSDA [10]	83.66	46.11	48.00	46.72
MTR [9]	86.08	51.68	40.81	40.14
WMTR [9]	81.89	46.25	51.29	47.87
BoF [11]	57.59	39.26	51.44	36.28
N-BoF [11]	62.70	42.28	61.41	41.63
B(TABL) [12]	73.62	66.16	68.81	67.12
C(TABL) [12]	78.01	72.03	74.04	72.84
DeepLOB[1]	78.91	78.47	78.91	77.66
Our model	80.73	80.62	80.73	**79.61**
Prediction Horizon k = 50				
RR [2]	43.90	43.60	43.30	42.70
SLFN [2]	47.30	46.80	46.40	45.90
BoF [11]	50.21	42.56	49.57	39.56
N-BoF [11]	56.52	47.20	58.17	46.15
B(TABL) [12]	69.54	69.12	68.84	68.84
C(TABL) [12]	74.81	74.58	74.27	74.32
DeepLOB[1]	75.01	75.10	75.01	74.96
Our model	77.93	78.14	77.93	**77.90**
Prediction Horizon k = 100				
RR [2]	42.90	42.90	42.90	41.60
SLFN [2]	47.70	45.30	43.20	41.00
BoF [11]	50.97	42.48	47.84	40.84
N-BoF [11]	56.43	47.27	54.99	46.86
B(TABL) [12]	69.31	68.95	69.41	68.86
C(TABL) [12]	74.07	73.51	73.80	73.52
DeepLOB[1]	76.66	76.77	76.66	76.58
Our model	81.23	81.29	81.23	**81.21**

圖 3 第一種測試方式與其 baseline 比較

Model	Acurracy %	Precision %	Recall %	F1 %
Prediction Horizon k = 10				
SVM [13]	-	39.62	44.92	35.88
MLP [13]	-	47.81	60.78	48.27
CNN-I [14]	-	50.98	65.54	55.21
LSTM [13]	-	60.77	75.92	66.33
CNN-II [15]	-	56.00	45.00	44.00
B(TABL) [12]	78.91	68.04	71.21	69.20
C(TABL) [12]	84.70	76.95	78.44	77.63
DeepLOB[1]	84.47	84.00	84.47	83.40
Our model	84.89	84.36	84.89	**84.02**
Prediction Horizon k = 20				
SVM [13]	-	45.08	47.77	43.20
MLP [13]	-	51.33	65.20	51.12
CNN-I [14]	-	54.79	67.38	59.17
LSTM [13]	-	59.60	70.52	62.37
CNN-II [15]	-	-	-	-
B(TABL) [12]	70.8	63.14	62.25	62.22
C(TABL) [12]	73.74	67.18	66.94	66.93
DeepLOB[1]	74.85	74.06	74.85	72.82
Our model	76.13	75.30	76.13	**74.59**
Prediction Horizon k = 50				
SVM [13]	-	46.05	60.30	49.42
MLP [13]	-	55.21	67.14	55.95
CNN-I [14]	-	55.58	67.12	59.44
LSTM [13]	-	60.03	68.58	61.43
CNN-II [15]	-	56.00	47.00	47.00
B(TABL) [12]	75.58	74.58	73.09	73.64
C(TABL) [12]	79.87	79.05	77.04	78.44
DeepLOB[1]	80.51	80.38	80.51	80.35
Our model	81.83	81.74	81.83	**81.73**

圖 4 第二種測試方式與其 baseline 比較

Models	Forward(ms)	Number of Parameters
DeepLOB[1]	2.334	60k
Our Model	0.434	31k

圖 5 與參考論文之推論延遲比較

5. 結論

本專題對於開源的資料集進行中價走向的預測策略，最終採取建置深度學習網路，運用 DeepLOB 結構[1] 對於股市數據結構的特徵解析與數學推演，結合了能夠降低延遲並且同時能夠強化在深度學習中捕捉重要特徵的神經網路，超越了原作者所建立模型之預測之四項指標，並且能夠在相同的硬體環境達到五分之一延遲。

在模型的優化過程中，我們參考了許多 ImageNet 架構，並融合其概念進入模型當中，由於我們的測資終究與圖像資料不盡相同，每個「像素」所代表的涵意更是與圖像資料有所差距，因此融入各種 ImageNet 時並不是只有將其架構直接套入我們的模型，而是更進一步根據原論文了解其核心概念，再將優化的關鍵以我們模型的架構模式加以導入，藉以提升我們模型的準確度，其中也充滿許多挑戰。

針對模型的表現，除了 Accuracy 的單純提升外，我們在 F1 score 能夠獲得提升，說明著並非只有對於各項數據的 label 能夠有正確的判斷，而是奠基於其上能夠預測出相比 True Negative 更為重要的 True Positive。對於亂度非常高，且在短周期的區間內並無規律性的數據而言，是建置演算法策略最大的目標。

本專題使用之資料集為標籤分布不均衡，因此在達成上述之目標又增添了困難度，而也因為如此我們對於訓練時所使用的 timestamp 數量也從 100 降為 10，與其補捉較大時間區段的資訊，反而更加著重於瞬態的走向，與此同時也獲得了訓練模型速度較快之優勢以及推論時間延遲之降低。

本專題使用影像處理相關演算法先對時序列資料捕捉其中重要的資訊後，再使用時間卷積網路去解析資訊其中乘載的股市走向。此種模式取代了使用記憶性的時序深度網路，減少單一網路需要的參數量，但依然能對於時間序列的特徵做出辨別並且做出標籤判斷。可以作為相關情境訊號處理的新的想法出發點。

心得感想

藉由實作專題研讀與自身研究領域相關的論文，也先行閱讀金融科技和機器學習相關文本，備足與自身研究領域相關背景知識，再透過每週匯報進度以及與實驗室學長討論，逐漸確認專題研究方向；在專題研究當中，也培養出自主安排進度、訂定規劃以及溝通討論的默契。

經過一個多學期的專題研究，我們從完全沒接觸過機器學習領域，到補足先備知識並進行機器學習的實作，中間花了不少時間與努力，也透過研究各種優化模型的方法，並套入到模型當中嘗試可行性，最終得出相比原先 baseline 更高的預測能力與更低的延遲。

對於本專題，我們認為依然有很大的空間可以延伸。首先是軟體於硬體的實現，演算法於邊緣裝置(edge device)的實現會因為現實情況而受到妥協，精準度和延遲是否能夠於系統上維持表現成為最大的困境。另外，演算法的通用受到開源資料透明度的限制，但金融資料的透明度逐日提升，因此如何能以個人單位制定捕捉訊號特徵和統整的機制將成為此領域的突破口。最終，我們在時序資料的處理上同時使用了影像處理與時序分析之網路結構，但計算負擔小且所需參數少之時序模型仍是訊號處理相關尚未突破方向，若能成功制定正確的演算法，將會帶給訊號處理領域和生活應用的一大突破。

參考資料

[1] Zhang, Z., Zohren, S., & Roberts, S. (2019). Deeplob: Deep convolutional neural networks for limit order books. IEEE Transactions on Signal Processing, 67(11), 3001-3012.

[2] Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. Journal of Forecasting, 37(8), 852-866.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[4] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[5] Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T. Y., ... & Zoph, B. (2021). Revisiting ResNets: Improved training and scaling strategies. Advances in Neural Information Processing Systems, 34, 22614-22627.

[6] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[7] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324).

[8] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[9] D. T. Tran, M. Magris, J. Kanniainen, M. Gabbouj,and A. Iosifidis, “Tensor representation in high-frequency financial data for price change prediction,” in Computational Intelligence (SSCI), 2017 IEEE Symposium Serieson. IEEE, 2017.

[10] D. T. Tran, M. Gabbouj, and A. Iosifidis, “Multilinear class-specific discriminant analysis,” Pattern Recognition Letters, vol. 100.

[11] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, “Temporal bag-of-features learning for predicting mid price movements using high frequency limit order book data,” IEEE Transactions on Emerging Topics in Computational Intelligence, 2018.

[12] D. T. Tran, A. Iosifidis, J. Kanniainen, and M. Gabbouj, “Temporal attention-augmented bilinear network for financial time-series data analysis,” IEEE transactions on neural networks and learning systems, 2018.

[13] “Using deep learning to detect price change indications in financial markets,” in Signal Processing Conference (EUSIPCO), 2017 25th European. IEEE, 2017.

[14] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, “Forecasting stock prices from the limit order book using convolutional neural networks,” in Business Informatics (CBI), 2017 IEEE 19th Conference on, vol. 1. IEEE, 2017.

[15] “Using Deep Learning for price prediction by exploiting stationary limit order book features,” arXiv preprint arXiv:1810.09965, 2018.