

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

指導教授：翁詠祿

組別: B253

組員: 陳昭維、林禹陞

Abstract

本篇專題以參考論文 “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)數據進行前處理與預測股市的走向。

透過參考論文以及對於金融時序列資料之專業知識，建立深度學習模型來做為預測股市的架構。藉由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。

Methodology

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 $x_t = [p_a^{(i)}, v_a^{(i)}, p_b^{(i)}, v_b^{(i)}]_{i=1}^{10}$ 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：

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

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：

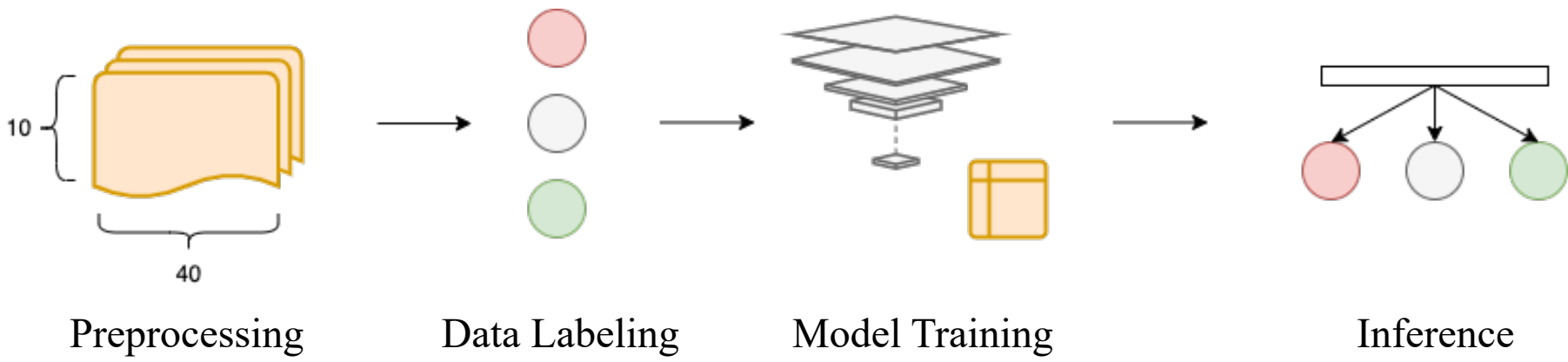
$$p_t = \frac{p_a^{(1)} + p_b^{(1)}}{2}$$

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：

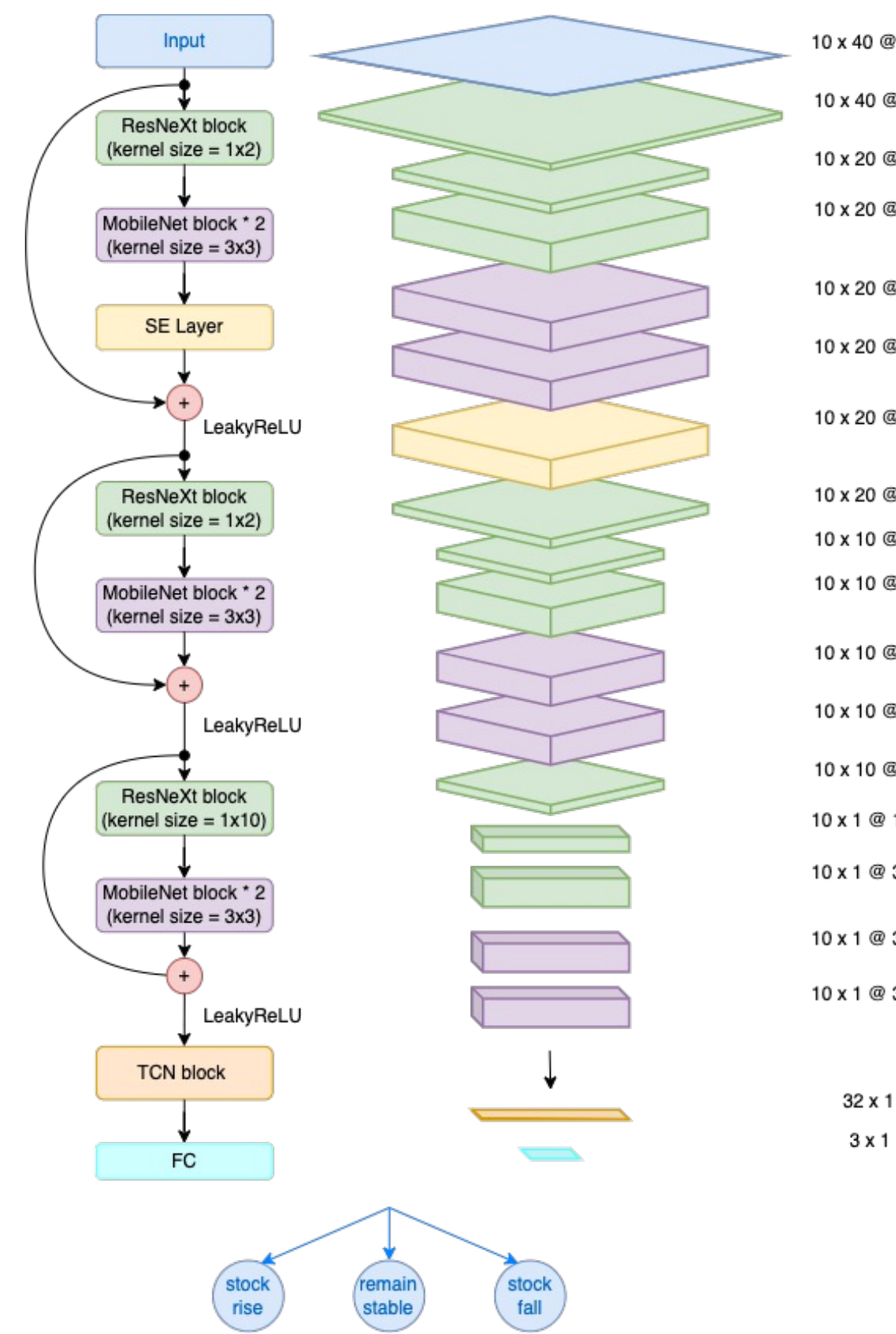
$$m_- = \frac{1}{k} \sum_{i=0}^k p_{t-i} \quad \text{and} \quad m_+ = \frac{1}{k} \sum_{i=0}^k p_{t+i}$$

作為t時刻時的漲跌方向，我們又訂一個閾值 α 做為判別漲跌的參數：

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



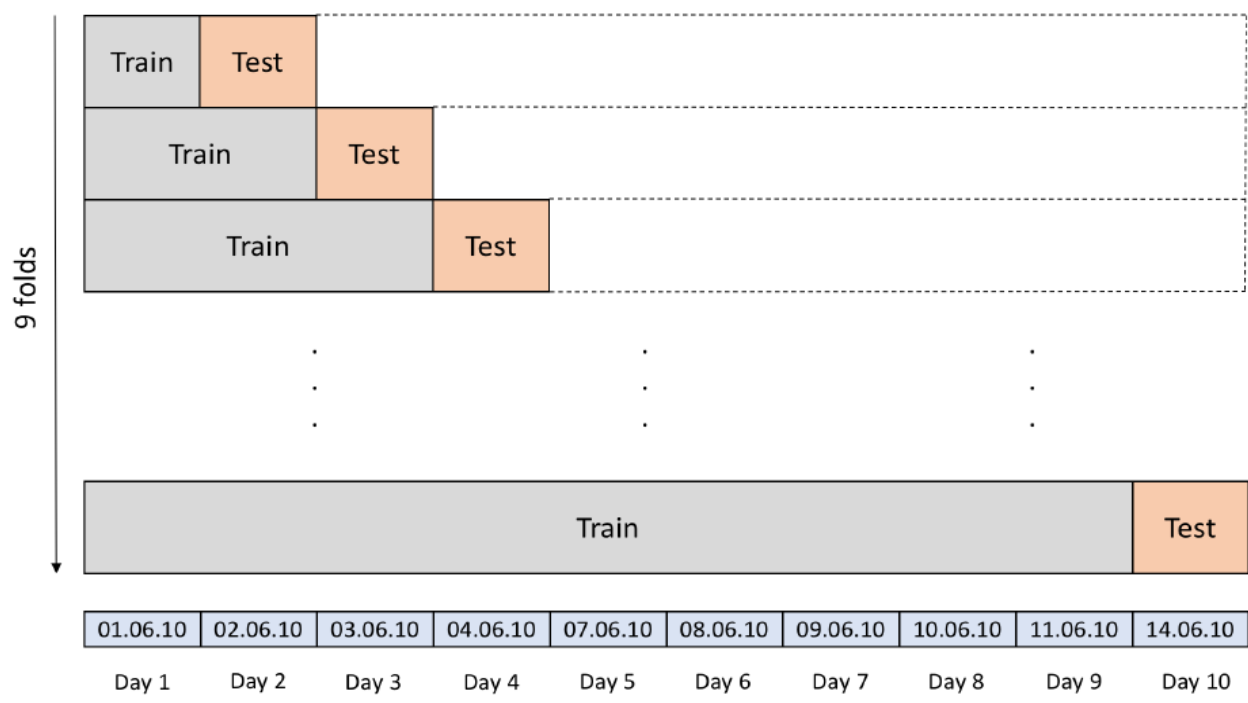
Model structure



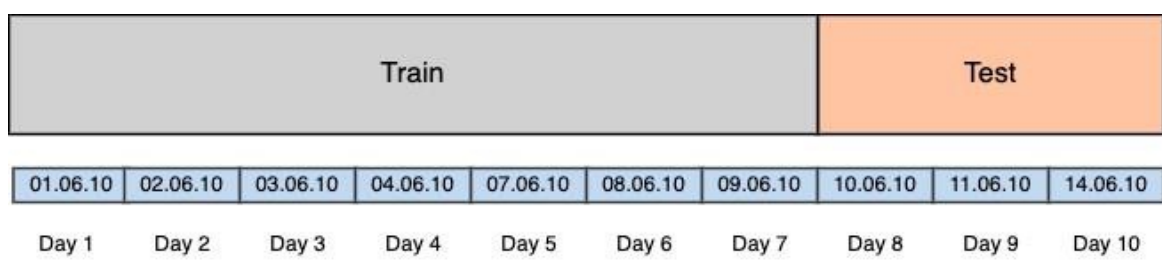
Test & Result

右表為我們的實作結果，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。

下圖為兩種測試方式，第一種使用前 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；第二種則為使用前 7 天的資料訓練後，去驗證後 3 天的資料。



Method 1



Method 2

Model	Accuracy %	Precision %	Recall %	F1 %
Prediction Horizon k = 10				
RR	48.00	41.80	43.50	41.00
SLFN	64.30	51.20	36.60	32.70
LDA	63.83	37.93	45.80	36.28
MDA	71.92	44.21	60.07	46.06
MCSDA	83.66	46.11	48.00	46.72
MTR	86.08	51.68	40.81	40.14
WMTR	81.89	46.25	51.29	47.87
BoF	57.59	39.26	51.44	36.28
N-BoF	62.70	42.28	61.41	41.63
B(TABL)	73.62	66.16	68.81	67.12
C(TABL)	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	43.90	43.60	43.30	42.70
SLFN	47.30	46.80	46.40	45.90
BoF	50.21	42.56	49.57	39.56
N-BoF	56.52	47.20	58.17	46.15
B(TABL)	69.54	69.12	68.84	68.84
C(TABL)	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	42.90	42.90	42.90	41.60
SLFN	47.70	45.30	43.20	41.00
BoF	50.97	42.48	47.84	40.84
N-BoF	56.43	47.27	54.99	46.86
B(TABL)	69.31	68.95	69.41	68.86
C(TABL)	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

第一種 method 與其他 baseline 之比較

Model	Accuracy %	Precision %	Recall %	F1 %
Prediction Horizon k = 10				
SVM	-	39.62	44.92	35.88
MLP	-	47.81	60.78	48.27
CNN-I	-	50.98	65.54	55.21
LSTM	-	60.77	75.92	66.33
CNN-II	-	56.00	45.00	44.00
B(TABL)	78.91	68.04	71.21	69.20
C(TABL)	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	-	45.08	47.77	43.20
MLP	-	51.33	65.20	51.12
CNN-I	-	54.79	67.38	59.17
LSTM	-	59.60	70.52	62.37
CNN-II	-	-	-	-
B(TABL)	70.8	63.14	62.25	62.22
C(TABL)	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	-	46.05	60.30	49.42
MLP	-	55.21	67.14	55.95
CNN-I	-	55.58	67.12	59.44
LSTM	-	60.03	68.58	61.43
CNN-II	-	56.00	47.00	47.00
B(TABL)	75.58	64.58	73.09	73.64
C(TABL)	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

第二種 method 與其他 baseline 之比較

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

與參考論文模型延遲&參數量比較

Conclusion

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。