基於情緒指標的 MF-SVM 模型——預測台積電每日報酬率趨勢

宋晉德

國立台北大學

摘要

所謂的投資，沒有做過任何量化回測分析，不了解各種數字的眉角，就像是在賭場中，閉著眼睛，玩吃餃子老虎。可是，如果，我們使用最新財務工程的技術——情緒指標；並且，結合機器學習，或許，這一切將有所不同，我們開始在意過往數據，以及數據中，所存在的意義。

因為方便研究，再加上台積電為台灣加權指數佔有很大比例，所以我們使用資料為，台積電股市交易價格；其資料長度為 2018 年 1 月 2 日至 2019 年 4 月 12 日。並且導入兩種情緒指標：VAR(關盤價的變異數)、AR(人氣指標)，以及一般指標:AVG(平均報酬率)；然後，使用改量過後的歸屬函數(MF)，來當作 SVM 訓練用的類別，進而提升測試表現。

關鍵字：SVM、情緒指標、歸屬函數(MF)、台積電

一、緒論

情緒指標應用於股票預測，是行為財務很熱門的議題；最主要原因，情緒指標可以代表市場投資者，目前對於市場是否有過度自信、或過度悲傷的跡象。通常，市場開始有較大的變化時，情緒指標也會有較大的波動的反應；最典型莫過於 VIX 指標(恐慌指標)，其是芝加哥期權交易所市場波動率指數的交易代號，通常用於衡量標準普爾 500 指數期權的隱含波動性。只是在這邊，我們是使用較為簡單的情緒指標：關盤價的變異數和人氣指標；人氣指標，是反應市場買賣的人氣。

SVM，Support Vector Machine，支援向量機；是機器學習主要基礎的模型，也是目前學術理論基礎最扎實的模型。SVM 相對於其他模型(像是 NN 或 DNN)，會比較重視維度的轉換，期許抽取出來的特徵(features)，轉換到某個維度上，會有一條線，可以把兩種不同類別切割出來(soft margin)。只是，在這邊我們是使用 linear regression，主要原因，是因為我們想導入歸屬函數，所以必須要有描述不同程度的類別。

歸屬函數，是 Fuzzy logic，作為模糊化使用的；可以幫助我們反應不同程度的市場氣氛。像是現在市場是很有朝氣的，strongly buy；或者是，市場穩定，buy；市場可能要注意了，代表 seldom buy；市場一攤死水，可以用 never buy 表示。

至於為什麼使用台積電股票，最主要原因，台積電是台灣加權指數，加權權重最大的標的，可以說是想了解台灣市場，就必須先了解台積電。

我們的 google colab 程式碼，已經放到 github 了。
(https://github.com/SquirrelMan/fuzzy_svm/blob/master/fuzzy.ipynb)

二、文獻回顧

「以情緒指標觀察台灣股市之研究」，在此專案中，探討了各種情緒指標，以及如何建立指標；並且導入單根檢定，去趨勢化，使 SVM 模型能趨於穩定。但是，在此專案中，並沒有使用到歸屬函數，以及為了方便我們檢驗歸屬函數的成效，我們使用的情緒指標類型會比較簡單。

「A New Fuzzy Support Vector Machine to Evaluate Credit Risk」，在此論文中，探討歸屬函數如何應用在 SVM，以及提供了另一種思維，同一筆資料，同時存在兩種對立面的類別。傳統 SVM 模型會是，[xi,yi]，但是，在此論文改寫成，[xi,yi=1,mk]和[xi,yi=-1,1-mk]，mk 為其類別對應的歸屬函數。但是，在此論文中，並沒有討論台積電股票的資料，而且其歸屬函數較為簡單；我們有改寫新的歸屬函數，進而應用我們想要用的資料。

三、數據

我們實作程式環境是使用 pycharm 和 google colab，它們編譯環境都是 python，透過 pycharm，取得數據資料；然後透過 google colab 實作 SVM 模型。

台積電(2330)數據，是使用 python 的 fix_yahoo_finance 數據庫，資料長度 2018 年 1 月 2 日至 2019 年 4 月 12 日。單根檢定範圍 2018 年 1 月 2 日至 2019 年 4 月 12 日。協整檢定範圍 2018 年 1 月 17 日至 2019 年 4 月 12 日。回測範圍 2018 年 1 月 19 日至 2019 年 4 月 12 日。會建立兩種情緒指標：關盤價變異數 VAR(長度為 11 天)、人氣指標 AR(長度為 11 天)；以及一般指標:平均報酬率 AVG(長度為 11 天)。資料測試類別，是使用報酬率大於(或等於零)為感興趣類別；報酬率小於零為不感興趣類別。

1.台積電數據、每日報酬率建立

如圖 1，是台積電每日的交易價格，而我們每日報酬率的建立是該日關盤價減去開盤價，然後除以開盤價，如公式 1:

```
stk_adj_ratio=(stk_close-stk_open)/stk_open (公式 1)
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2018-01-02	231.5	232.5	231.0	232.5	224.324173	18055269
1	2018-01-03	236.0	238.0	235.5	237.0	228.665939	29308091
2	2018-01-04	240.0	240.0	236.5	239.5	231.078018	29096613
3	2018-01-05	240.0	240.0	238.0	240.0	231.560440	22438255
4	2018-01-08	242.0	242.5	240.5	242.0	233.490112	20233692

	Date	Open	High	Low	Close	Adj Close	Volume
302	2019-04-08	251.0	253.0	250.5	253.0	253.0	45184821
303	2019-04-09	253.0	254.0	252.0	254.0	254.0	22355674
304	2019-04-10	253.0	254.5	252.0	254.0	254.0	25849934
305	2019-04-11	253.0	254.0	251.5	252.0	252.0	24896840
306	2019-04-12	251.5	253.0	251.0	252.0	252.0	13548148

圖 1、台積電每日的 OHLCV

並且，對台積電的每日報酬率，進行單根檢定觀察其是否去趨勢，如圖 2，可以發現 p-value 遠小於 0.05，代表資料已經趨於穩定(去趨勢)，如圖 3，數值大約介於正負 0.03 之間。

	value
Test Statistic Value	-19.7067
p-value	0
Lags Used	0
Number of Observations Used	306
Critical Value(1%)	-3.4519
Critical Value(5%)	-2.87103
Critical Value(10%)	-2.57183

圖 2、台積電的每日報酬率、p-value 遠小於 0.05、lags=0

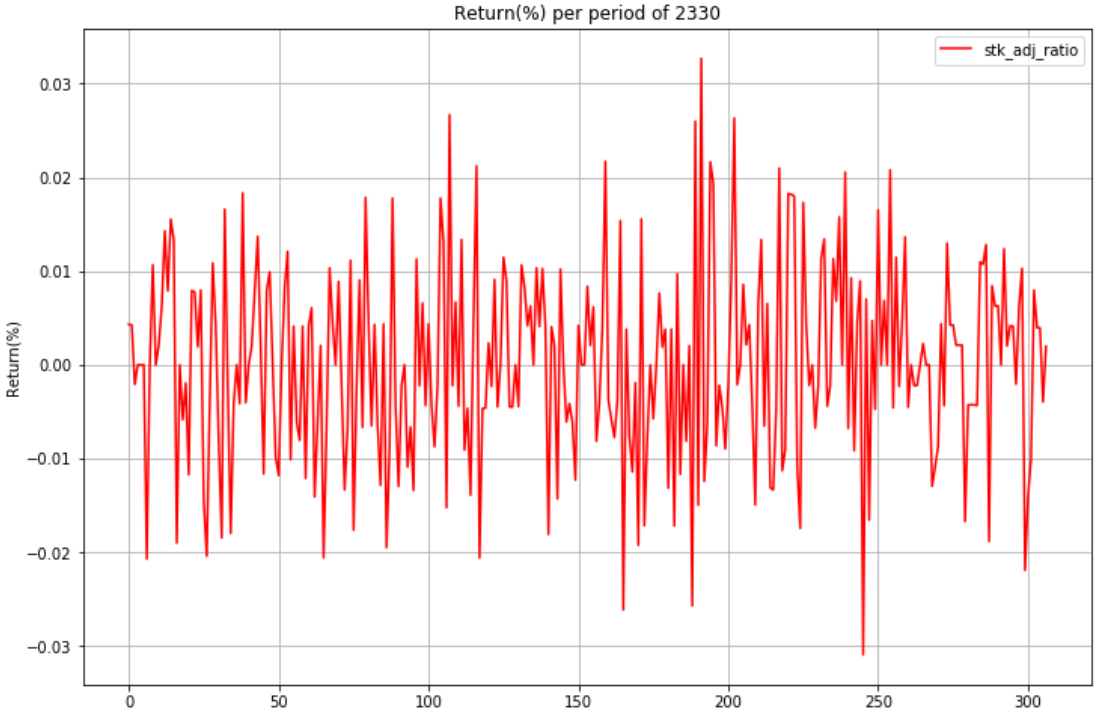

圖 3、台積電的每日報酬率，去趨勢化

2.關盤價變異數 VAR(長度為 11 天)

如公式 2，是我們的 VAR 指標建立，取得前 11 天(包含該日)的關盤價，然後計算變異數，當作該日的 VAR 指標。

```
stk_adj_var_11.append(np.var(stk_close[i-11:i]))(公式 2)
```

並且，對 VAR，進行單根檢定觀察其是否去趨勢，如圖 4，可以發現 p-value 遠小於 0.05(SVM 模型，資料會使用到 lag=1 和 lag=2)，代表資料已經趨於穩定(去趨勢)，如圖 5，數值大約介於 0 到 140 之間。VAR 對於每日報酬率，協整檢定遠小於 0.05，如圖 6。

	value
Test Statistic Value	-7.43278
p-value	6.29659e-11
Lags Used	2
Number of Observations Used	293
Critical Value(1%)	-3.45287
Critical Value(5%)	-2.87146
Critical Value(10%)	-2.57205

圖 4、VAR，單根檢定，p-value 遠小於 0.05、lags=2

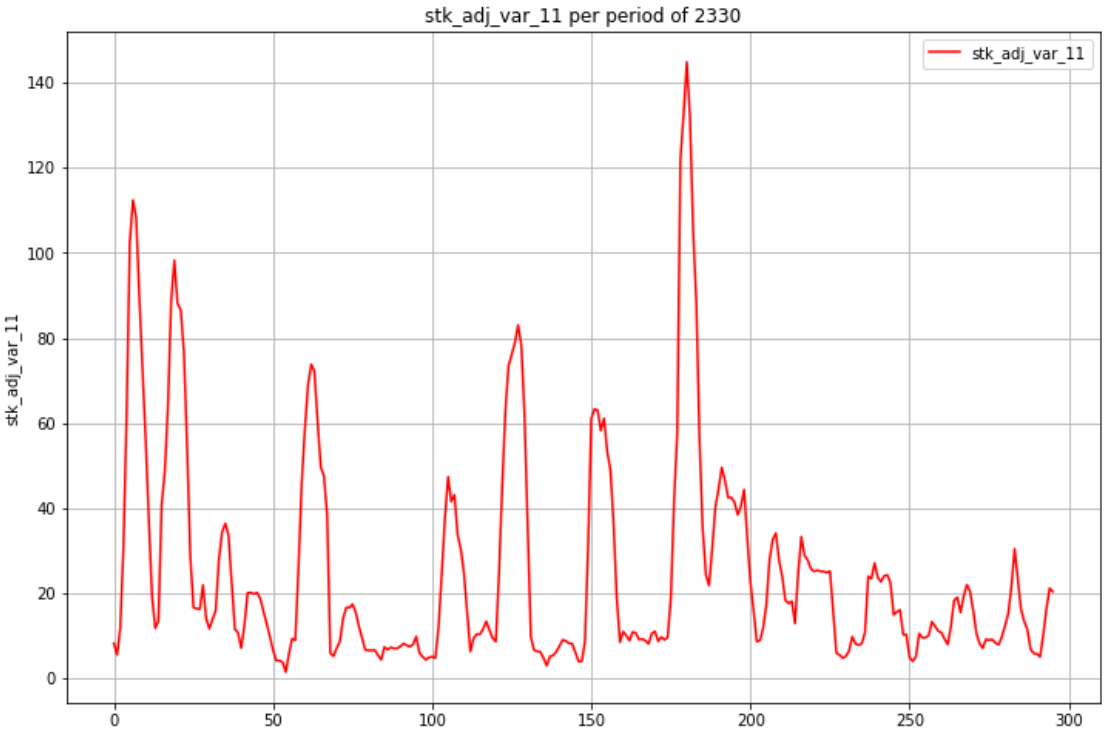

圖 5、VAR，去趨勢化

```
result = sm.tsa.stattools.coint(stk_adj_ratio[11:],stk_adj_var_11)
pvalue = result[1]
pvalue
```

```
0.0
```

圖 6、VAR，協整檢定，p-value 遠小於 0.05

3.人氣指標 AR(長度為 11 天)

如公式 3，是我們的 AR 指標建立，分子為前 11 天(包含該日)最高價減去開盤價的合計，然後分母為前 11 天(包含該日)開盤價減去最低價的合計，當作該日的 AR 指標。

```
stk_adj_ar_11.append((np.sum(stk_high[i-11:i])-np.sum(stk_open[i-11:i]))/(np.sum(stk_open[i-11:i])-np.sum(stk_low[i-11:i])))
```
(公式 3)

並且，對 AR，進行單根檢定觀察其是否去趨勢，如圖 7，可以發現 p-value 遠小於 0.05，代表資料已經趨於穩定(去趨勢)，如圖 8，數值大約介於 0 到 2.5 之間。VAR 對於每日報酬率，協整檢定遠小於 0.05，如圖 9。

	value
Test Statistic Value	-3.09656
p-value	0.0268088
Lags Used	11
Number of Observations Used	284
Critical Value(1%)	-3.45359
Critical Value(5%)	-2.87177
Critical Value(10%)	-2.57222

圖 7、AR，單根檢定，p-value 遠小於 0.05、lags=11

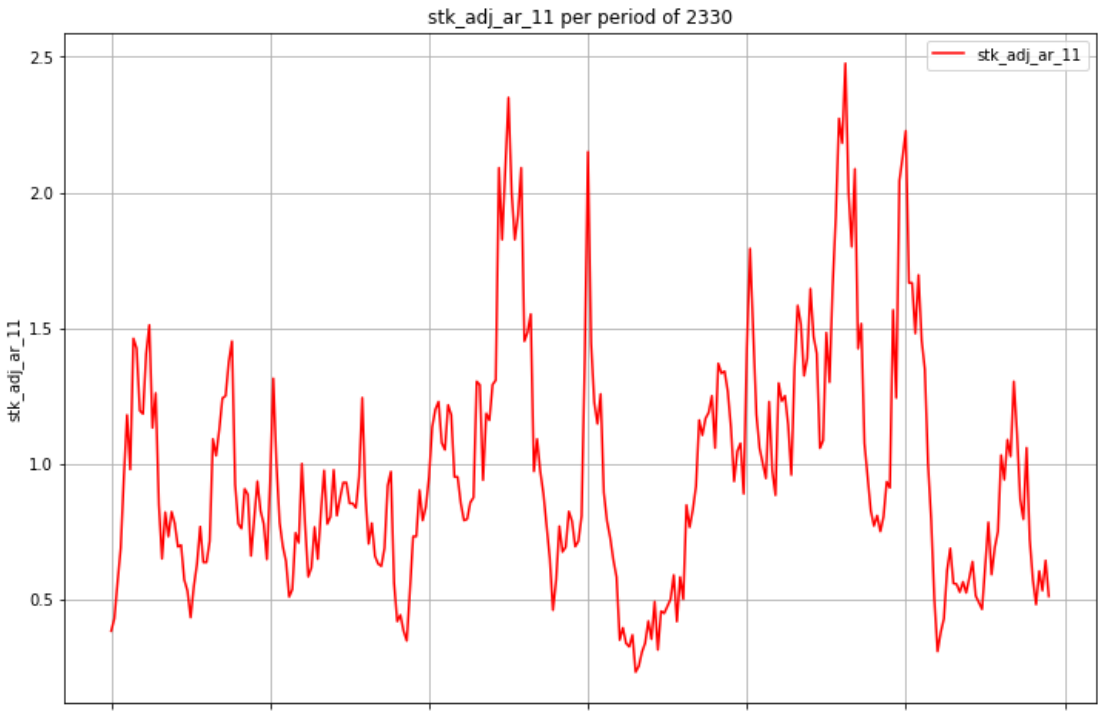

圖 8、AR，去趨勢化

```
result = sm.tsa.stattools.coint(stk_adj_ratio[11:],stk_adj_ar_11)
pvalue = result[1]
pvalue
```

```
0.0
```

圖 9、AR，協整檢定，p-value 遠小於 0.05

4.平均報酬率 AVG (長度為 11 天)

如公式 4，是我們的 AVG 指標建立，為前 11 天(包含該日)的平均報酬率，當作該日的 AVG 指標。

```
stk_adj_avg_11.append(np.mean(stk_adj_ratio[i-11:i]))
 (公式 4)
```

並且，對 AVG，進行單根檢定觀察其是否去趨勢，如圖 10，可以發現 p-value 遠小於 0.05，代表資料已經趨於穩定(去趨勢)，如圖 11，數值大約介於-0.006 到 0.006 之間。AVG 對於每日報酬率，協整檢定遠小於 0.05，如圖 12。

	value
Test Statistic Value	-3.19972
p-value	0.0200011
Lags Used	15
Number of Observations Used	280
Critical Value(1%)	-3.45392
Critical Value(5%)	-2.87192
Critical Value(10%)	-2.5723

圖 10、AVG，單根檢定，p-value 遠小於 0.05、lags=15

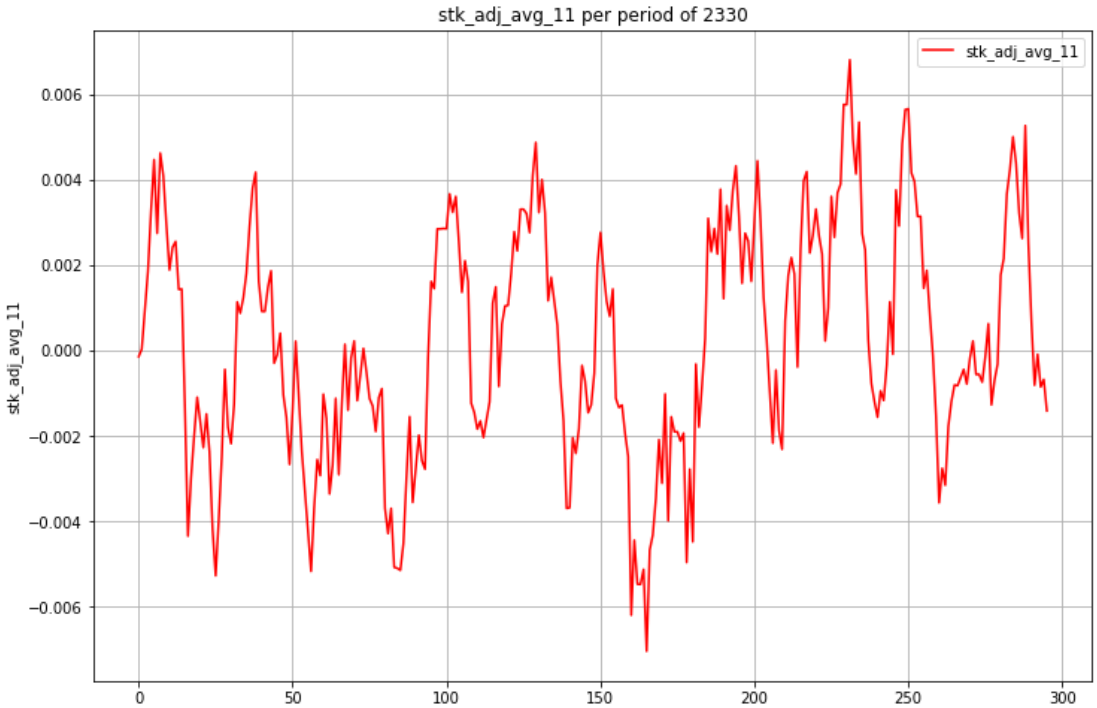

圖 11、AVG，去趨勢化

```
result = sm.tsa.stattools.coint(stk_adj_ratio[11:],stk_adj_avg_11)
pvalue = result[1]
pvalue
```

```
0.0
```

圖 12、AVG，協整檢定，p-value 遠小於 0.05

四、實驗方法

1.建立每筆資料的 features

如圖 13，可以看到我每筆資料的 features。lag1，代表資料往後推移 1 筆，lag2，代表資料往後推移 2 筆。

	stk_adj_ratio_lag1	stk_adj_var_11	stk_adj_var_11_lag1	stk_adj_var_11_lag2	stk_adj_ar_11	stk_adj_avg_11
2	1.428571e-02	11.834711	5.561983	8.231405	0.562500	0.000953
3	7.889546e-03	30.743802	11.834711	5.561983	0.688889	0.001859
4	1.553398e-02	62.194215	30.743802	11.834711	0.952381	0.003272
5	1.333333e-02	102.458678	62.194215	30.743802	1.179487	0.004484
6	-1.901141e-02	112.376033	102.458678	62.194215	0.978723	0.002755
7	1.387234e-18	108.289256	112.376033	102.458678	1.461538	0.004637
8	-5.847953e-03	89.446281	108.289256	112.376033	1.425000	0.004106
9	-1.930502e-03	72.566116	89.446281	108.289256	1.195652	0.002961
10	-1.171875e-02	55.698347	72.566116	89.446281	1.183673	0.001896
11	7.905138e-03	36.884298	55.698347	72.566116	1.404255	0.002425
12	7.766990e-03	19.095041	36.884298	55.698347	1.511111	0.002564
13	1.930502e-03	11.822314	19.095041	36.884298	1.132075	0.001441
14	7.968127e-03	13.537190	11.822314	19.095041	1.260000	0.001448

圖 13、每筆資料的 features：為了配合 lag1 和 lag2，前面兩筆資料(第 0 筆和第 1 筆)，就直接丟掉

2.一般標籤化方式(測試資料)

如公式 5，假設股價盤整以上，我們就是持有台積電，所以每日報酬率為正(或等於零時)，我們給予標籤是 1；為負時，我們給予標籤是 0。

```
label.append((np.array(all_stk_adj_ratio)[i][0]>=0)*1)(公式 5)
```

3.透過歸屬函數(membership function;MF)進行標籤化(訓練資料)

如公式 6，我們改良 Yongqiao Wang, Shouyang Wang, and K. K. Lai 歸屬函數定義。我們使用原本圖 14 的 Probit，進行改良，「買入台積電(buy)」外面改成 exp，當獲利越多時，其標籤化加權要越大(代表 strongly buy 的程度)；越靠近 1 時，代表只有獲利一點點(代表 buy)；小於 1 時，代表沒有獲利(代表 seldom buy)；越接近 0 時，代表賠很多(代表 never buy)。相反的，「賣出台積電(sell)」，外面改成 exp 之後，對於 Y 軸截距我們往上移動 0.5，加強「賣出台積電(sell)」的模糊空間，並且最後再開根號，降低一點點偵測的敏感度。在 Yongqiao Wang, Shouyang Wang, and K. K. Lai 的論文裡面，有提到每筆資料都會有兩個標籤和其對應的歸屬函數；在這邊我們代表的是，「買入台積電(buy)」和「賣出台積電(sell)」，並且每筆資料都會得到兩種歸屬函數的數值。但是，為了之後訓練資料方便，我們有把這兩個類別進行合併，「買入台積電(buy)」的歸屬函數大於(或等於)「賣出台積電(sell)」，其類別的歸屬函數就會是「買入台積電(buy)」；但是，相反地，就會是「買入台積電(buy)」的歸屬函數的一半，強化 seldom buy、never buy(以及少部分 buy)的模糊空間。

```
temp_buy = math.exp(((np.array(all_stk_adj_ratio)[i][0])-0.000115)
/0.010516)
temp_sell =math.sqrt(math.exp(((np.array(all_stk_adj_ratio)[i][0])
-0.000115)/0.010516*(-1))+0.5)
if temp_buy >= temp_sell:
  label.append(temp_buy)
else:
  label.append(temp_buy/2) (公式 6)
```

$$\text{Linear}: m_{ks} = \frac{s_k - \min\limits_{k=1,\ldots,N} s_k}{\max\limits_{k=1,\ldots,N} s_k - \min\limits_{k=1,\ldots,N} s_k}$$

$$\text{Bridge}: m_{ks} = \begin{cases} 1 & s_k > \overline{s} \\ \frac{s_k - \underline{s}}{\overline{s} - \underline{s}} & \underline{s} < s_k \leq \overline{s} \\ 0 & s_k \leq \underline{s} \end{cases}$$

$$\text{Logistic}: m_{ks} = \frac{a^{as_k+b}}{a^{as_k+b}+1}$$

$$\text{Probit}: m_{ks} = \Phi\left(\frac{s_k - \mu}{\sigma}\right)$$

圖 14、Yongqiao Wang, Shouyang Wang, and K. K. Lai 歸屬函數定義

4.訓練資料和測試資料

回測資料，總共有 293 筆(日)。我們訓練資料使用前面 193 筆，測試資料使用後面 100 筆。訓練資料和測試資料，進行不同標籤化的方式，主要原因是因為想強調訓練資料的，獲利越高時，其加權應該要越高；當在進行測試資料時，獲利的類別(感興趣類別)，其偵測效果會越好，很適合做多頭交易策略。

5.SVM 模型使用 linear regression，並使用混淆矩陣作為測試表現的依據，如公式 7

```
confusion_matrix = [[count_TP,count_FN],[count_FP,count_TN]](公式 7)
```

四、實驗數據

從表 1，可以看到，透過添加歸屬函數，可以使原來的 SVM 的模型，在 Precision 的部分提升 2%；甚至在 Recall 的部分，提升了 31%。另外，從圖 15 和圖 16，可

以知道，在 100 天交易內，MF-SVM 整體獲利比 SVM 提升約 2%左右。從實驗數據，可以了解，透過添加歸屬函數，能幫助我們改善交易績效。

	SVM	MF-SVM
TP	24	43
FN	36	17
FP	14	23
TN	26	17
Precision(TP/(TP+FP))	0.63157894736	0.65151515151
Recall(TP/(TP+FN))	0.4	0.71666666666

表 1、SVM 和 MF-SVM 模型表現比較

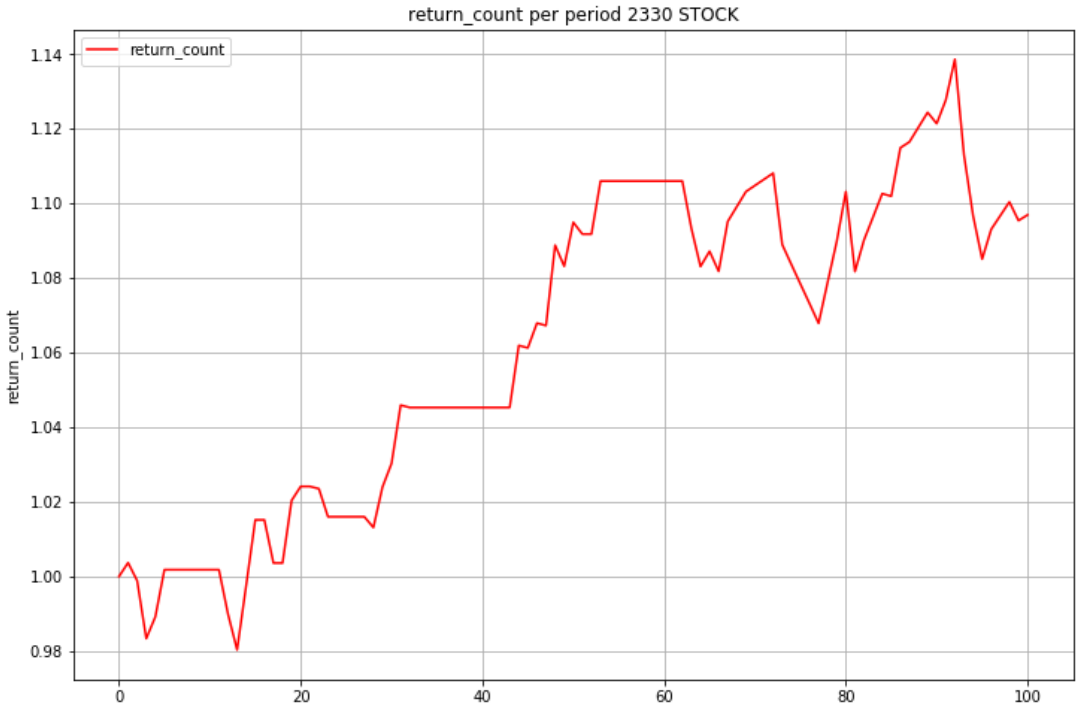

圖 15、MF-SVM，多頭交易當日沖策略，滑價 0.9994

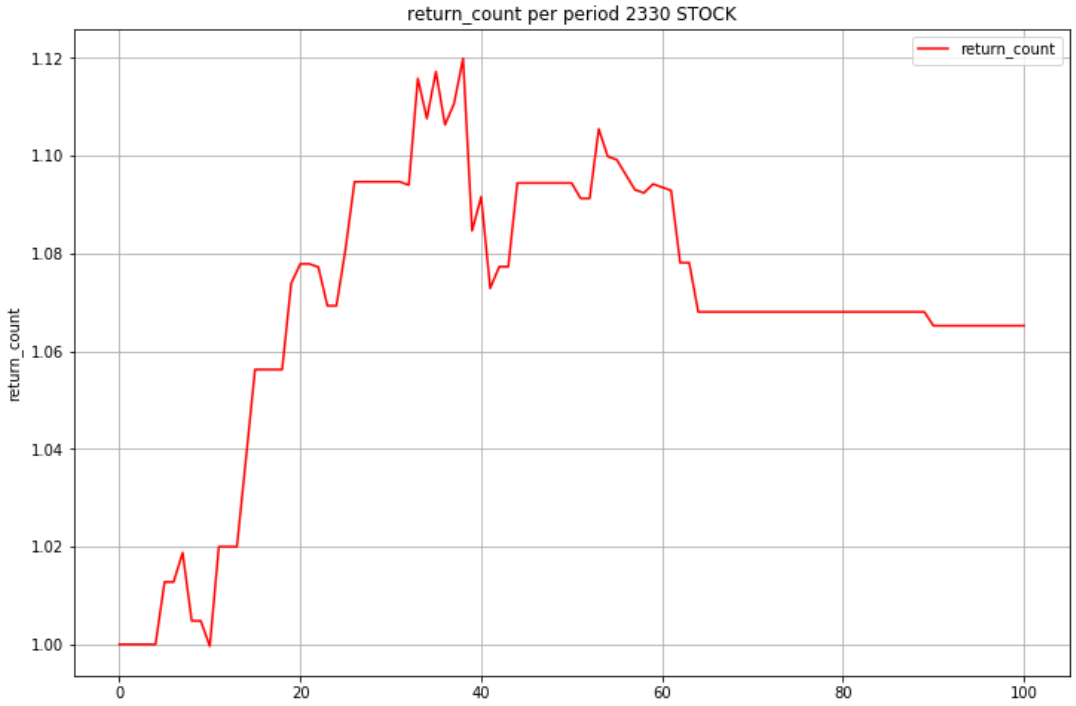

圖 16、SVM，多頭交易當日沖策略，滑價 0.9994

五、結論與建議

在台積電交易的獲利表現上，可以發現到我們改良過後的 MF-SVM 是有提升的。或許，下次我們可以添加更多情緒指標的 features，並且結合 information gain 的篩選，來幫助提升 MF-SVM 的績效表現。

參考文獻

1. 宋晉德、陳映如、黃柏毅等等，「以情緒指標觀察台灣股市之研究」，資產管理碩士課程，證基會，2019。(https://github.com/SquirrelMan/svm_sentiment)
2. Yongqiao Wang, Shouyang Wang, and K. K. Lai,"A New Fuzzy Support Vector Machine to Evaluate Credit Risk", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 13, NO. 6, DECEMBER 200

Song Jin-De Graduate Student , Master of Computer Science and Information Engineering in National Taipei University

Github:https://github.com/SquirrelMan

Linkedin:
https://www.linkedin.com/in/jim-song-90506213a

Gmail:jim845192000@gmail.com