中華大學資訊工程學系

104 學年度專題製作期末報告

股票分析 APP 之開發

指導老師：梁秋國教授

專案組員：B10102106 薛賀聰

中華民國一百零五年六月

目錄

第一章　序論

1.1 研究動機與目標

股市是眾多投資方法中最能獲取高額報酬的方法，但同時它也是眾多投資方法中最具風險的一種。投資者藉由投資股票市場享有企業股份的獲利及盈餘分配的權利，企業也藉由股市取得資金的需求。故股票市場對於投資者、企業及整體社會有很大的經濟貢獻。

現在的股票 app 有很多，看股票價位、看歷史線圖的，能直接下單的等等，而我想法則是大部份在投資股票的人士中，很多都是辦公室的上班族，因為工作的關係，並沒有時間長期在電腦或手機上監看股票市場的走勢，並且對股票市場作出一定的分析，於是一個能為使用者在某特定的時間，對使用者推送通知，為使用者分析股票漲跌的機率的 APP 就成為了我專題的方向。

而資料探勘是近年來被廣泛應用於各領域來協助分析大量資料的技術，主要的訴求是希望透過資料探勘的技術從大量的歷史資料中挖掘其中隱含有價值的知識。於是利用決策樹的方式，來分析某特定股票在過去半年的走勢，並利用分析所得的結果，預測該股票在未來一天漲跌的機率，並在 iPad 上顯示出來成為了這次專題的目標。

第二章　開發環境

2.1 開發平台及工具

2.1.1 Digital Ocean

Digital Ocean 是提供虛擬專用伺服器(VPS)的雲基礎架構提供商，有多種方案可以選擇，最便宜的方案為每個月 5 美元，提供 512 MB 記憶體、20GB SSD、單核心 CPU 和每月 1000GB 的網路流量，這也是這個專題中所使用的方案，系統為 CentOS 7.0 x64。

2.1.2 PyCharm

PyCharm 是由 JetBrains 開發的一款 Python IDE，帶有一整套可以幫助開發者在使用 Python 語言開發時提高其效率的工具，例如調試、Project 管理、代碼跳轉等。同時 PyCharm 也支援 Google App Engine 和 IronPython。

2.1.3 Xcode

Xcode 是蘋果公司向開發人員提供的整合式開發環境，用於開發 Mac OS X、iOS、WatchOS 和 tvOS 的應用程式。

2.2 使用語言及硬體設備

使用語言

- Python 2.3
- Swift
- MySQL
- JavaScript
- PHP

硬體設備

- Hackintosh laptop(Dell XPS 9550)
- Apple iPad Air 2
- Apple iPod touch 5

第三章　研究方法

3.1 研究理念

決策樹 ID3 為最早使用的決策樹演算法之一，ID3 的主要觀念是以遞迴的方式將訓練資料作切割。在每一次產生節點時，某些輸入的訓練子集將取出測試，以資訊獲取量來當作測試，在選取過後，將以具有最大資訊獲取量的值當作分支的節點，接下來依照其遞迴的動作選取下一個分支節點，直到每一個訓練資料都屬於一個分類之中 或是符合某個滿足條件。

C4.5 是 ID3 的延伸方法，改善了 ID3 產生過多子集合，而每個子集合僅包含少數資料的問題，並且具備處理連續數值型屬性、雜訊的處理，另外也兼具修剪樹的能力。C4.5 在決策樹的每個節點上使用資訊獲取量(Information Gain)來選擇測試屬性，選擇具有最高資訊獲取量(或最大 entropy 壓縮)的屬性作為當前節點的測試屬性。在解決分類的問題時，必須先給定一組資料集，稱為訓練資料集(training set)，資料集中每一筆資料都包含著數個屬性(attribute)及一個目標類別(target class)。C4.5 是採用由上而下（top-down），以資料為導向的分類法處理。

C5.0 是延續 C4.5 的演算法架構，提供了在許多應用上較受歡迎的規則集(rule sets)，將分類的條件以規則形式表達，增加了閱讀分類規則的可讀性。C5.0 可用來處理數值性(numeric)或是名目性(nominal)欄位的資料，分析結果可用容易理解的決策樹(decision trees)。C5.0 與 C4.5 不同之處在於 C5.0 可以處理更多種資料型態(data types)，如日期(date)、時間(times)、時間戳記(timestamps)、序列性的離散型資料(ordered discrete attributes)等資料型態，因此本專題將使用 C5.0 來產生決策樹並且預測結果。

3.2 研究步驟

3.2.1 股票篩選

基於伺服器的效能，要分析並預測臺灣股市上的一萬支以上的股票需要的時間和代價太高，幾乎是不可能。而且大部份的投資人士也只會對較大的股票分析感興趣，因此在這次專題中，只會集中分析 16 支股票。

2353 宏碁	2498 宏達電	3596 智易	2448 晶電
2303 聯電	0050 台灣 50	2311 日月光	2891 中信金
2317 鴻海	2002 中鋼	2881 富邦金	3481 群創
2330 台積電	3474 華亞科	2882 國泰金	2345 智邦

3.2.2 資料前處理

以 Linux 指令 crontab 工作排程的方式，利用 python script 每三秒鐘對臺灣證券交易所進行網路爬蟲，取得股票資料。並對取得的股票資料以日期、時間、代號進行分類，並產生或更新 csv 檔案。

```
'''
This method sends a GET request to TWSE with an UNIX timestamp and gets the real-time stock data
args dict :
- quote : a stock number with the abbreviation of the market (string)
'''
def get_stock_info(quote):
    req = requests.session()
    req.get('http://mis.twse.com.tw/stock/index.jsp',
            headers={'Accept-Language': 'zh-TW,zh;q=0.8,en-US;q=0.6,en;q=0.4'})
    response = req.get('http://mis.twse.com.tw/stock/api/getStockInfo.jsp?ex_ch={}'.format(quote))
    return response.text
```

利用 csv 檔的資料，計算簡單移動平均(Simple Moving Average, SMA)、成交量移動平均(Volume Moving Average, VMA)、KD 指標(Stochastic Oscillator, K, D)等資料，並以 Javascript 繪製股票圖表(K 線圖、股量圖、股價圖)。並在股市收盤後，產生當天各股票收盤資料及台股上市公司列表的 csv 檔。

```
'''
This method send a POST request to TWSE and gets the details daily closing of a specified stock
args dict :
- download : the file type of the return data, in this case, it is csv (string)
- selectType : the sector of the stock (string)
- qdate : the date of the data (int)
'''
def post_request(params):
    request = requests.post('http://www.twse.com.tw/ch/trading/exchange/MI_INDEX/MI_INDEX.php',
        params={'download': 'csv', 'selectType': selectType, 'qdate': date})
    # return nothing if there is an error
    if request.status_code == 200:
        return request.text
    else:
        write_err_log(request.status_code.__str__() + '    ' + request.reason)
        return ''
```

簡單移動平均線及 KD 指標的計算方式：

$$\text{SMA[t]} = \sum_{t}^{t-N+1} P[t] \ / \ \text{N}$$

$$\text{RSV[t]} = \frac{P[t]-lowest_9}{highest_9-lowest_9} \times 100$$

$$\text{K[t]} = \frac{2}{3} \times \text{K[t} - 1] + \frac{1}{3} \times RSV[t]$$

$$\text{D[t]} = \frac{2}{3} \times D[t-1] + \frac{1}{3} \times K[t]$$

其中 t：日期，P：收盤價，N：N 日，K：K 值，D：D 值，RSV：未成熟隨機值，$lowest_9$：最近 9 天最低價，$highest_9$：最近 9 天最高價

處理前的股票資料樣式：

最佳五檔買入數量	161_682_632_1090_1452_
最價五檔賣出數量	366_631_972_821_932_
日期	20160607
最佳五檔買入價格	80.40_80.30_80.20_80.10_80.00_
股票代號	2317
最佳五檔賣出價格	80.50_80.60_80.70_80.80_80.90_
當日累計成交量	25596
時間	13:30:00
當盤成交量	1890
最近成交價	80.50

處理後的股票資料樣式：

時間	13:30:00
當盤成交量	1890
最近成交價	80.50
SMA5	79.36
SMA20	79.09
SMA60	79.37
SMA120	79.13
SMA240	83.41

處理前的收盤資料樣式：

股票代號	2317
股票公司全名	鴻海精密工業股份有限公司
證券名稱	鴻海
成交股數	25,729,544
成交筆數	10,938
成交金額	2,069,313,707
開盤價	80.40
最高價	80.80
最低價	80.10
收盤價	80.50
漲跌(+/-)	+
漲跌價差	1.20
最後揭示買價	80.20
最後揭示買量	366
最後揭示賣價	80.50
最後揭示賣量	161
本益比	8.74

處理後的收盤資料樣式：

證券名稱	鴻海
股票公司全名	鴻海精密工業股份有限公司
開盤價	80.40
最高價	80.80
最低價	80.10
收盤價	80.50
成交筆數	10,938
VMA5	31709
VMA10	27196
K	38.71
D	49.20

3.2.3 決策樹產生方法

各個技術指標應只在某特定時間適用於某特定的股票，因此分析技術指標的方法應根據時間而改變，而且每支股票的習性會因著各種原因都有所不同，例如產業、公司營運目標、公司營運成果等，所以每支股票都應獨立處理。

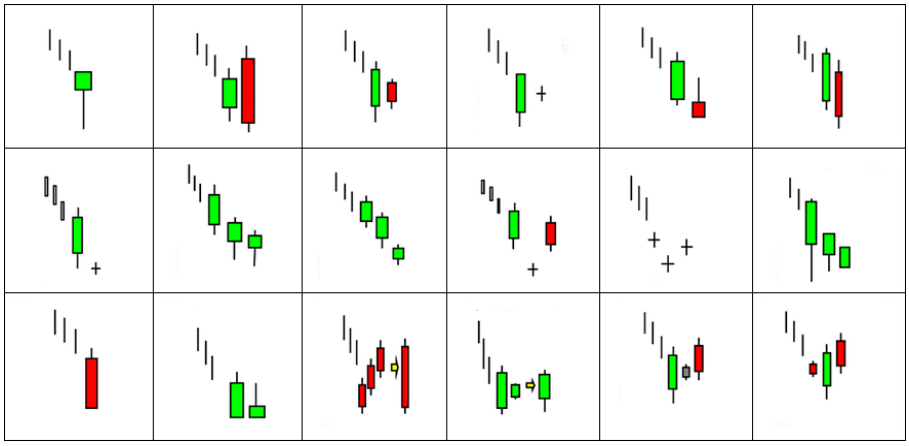
決策樹的其中一個優點是可以透過閱讀決策樹(seeing the decision tree)而理解其模型，但由於這專題是以每天預測一次的方式來執行，因此會產生很多過於複雜、人類並不可讀的模型。決策樹的訓練資料集為以當天起計算前半年的股票資料，為了避免出錯，測試樣本集(Testing set)設定為當天前一天的股價。

決策樹會以以下步驟及規則(不順序、部份)產生：

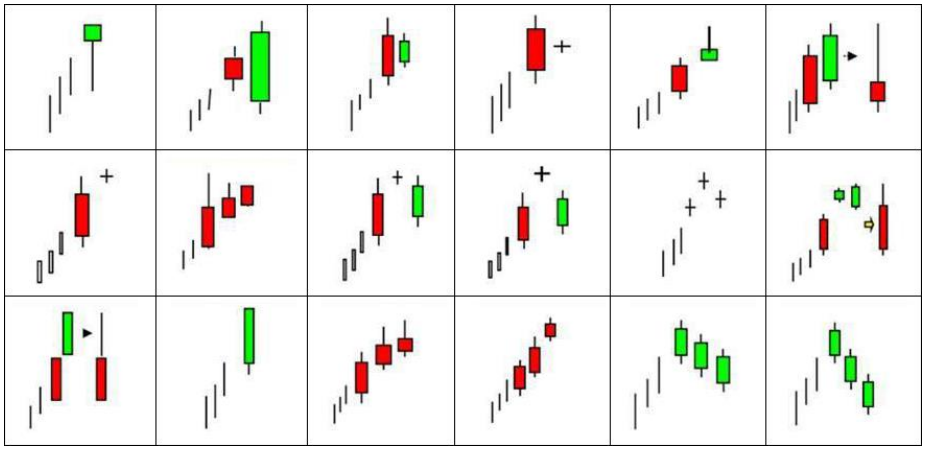
for each month in backtest：
    for each open day in the months：
        for each stock that has been selected：
            for each parameter value of the tree：
                for each fold of the cross-validation：
                    build a decision and evaluate it

1. 股票現價向上穿越 MA 為股價上漲，向下穿越 MA 為股價下跌。
2. 短期 MA 向上穿越長期 MA 為股價上漲，反之短期 MA 向下穿越長期 MA 為股價下跌。
3. 股票現價、短期 MA、中期 MA、長期 MA 依序由上而下排列為股價上漲，反之股票現價、短期 MA、中期 MA、長期 MA 依序由下而上排列，為股價下跌。
4. 短期 MA 向上穿越長期 MA 為股價上漲，反之短期 MA 向下穿越長期 MA 為股價下跌。
5. K 值由下往上穿越 D 值為股價上漲，由上往下穿越 D 值為股價下跌。
6. KD 值大於 80 時為股價下跌，KD 值小於 20 時為股價上升。
7. K 值連續 3 天以上大於 80 為股價上漲，K 值連續 3 天以上小於 20 為股價下跌。
8. 股價創波段新高時，KD 值沒有跟隨著創新高為股價下跌。股價創波段新低時，KD 值沒有跟隨著創新低為股價上漲。

9. 找出最近一個月的 K 線圖中兩個最低和最高點，以最低及最高兩點計算出兩個斜率，並利用斜率找出目前該線到達點之股價，若最低兩點的斜率所產生的股價、最高兩點的斜率所產生的股價低於目前股價，並且目前的成交量為上漲，則為股價上漲。
10. 分析 K 線圖，若出現以下圖表中的圖形，為股價上漲。



11. 分析 K 線圖，若出現以下圖表中的圖形，為股價下跌。



3.2.4參數調整

為了增加決策樹模型的預測能力，參數調整是重要的一環。本專題中所考量的參數為：

1. 決策樹節點的數量為多少
2. 哪一些節點的預測能力較高
3. 股票本身的行為(Behaves)如何
4. 股市資料會不會過於複雜而無法分類
5. 預測結果為長期或短期
6. 訓練資料集的大小對預測結果的影響

第四章　結論

4.1 結論

本專題所使用的決策樹主要是由各個技術指標作分類變項，例如 SMA、股價、KD 值、K 線圖類型等，找出股價變動的線索，模型建構的預測能力就成為關鍵因素，因為模型的判斷策略會決定預測準確率。且由實證結果，由當天起計算半年的股市數據作訓練資料集所建立之決策樹模型，藉由其產生的漲跌參考規則進而預測前一天的股價漲跌機率，其最終所獲得之平均預測準確率為 47%。在此足已見證，透過決策樹模型所得出的預測結果，有一定程度上的成效，值得進一步加以運用。

4.2 改善

本專題中平均預測準確率為 47%，其原因為本專題只集中於技術分析層面，技術分析主要是根據過去的股價資料，嘗試分析過去的趨勢，以預測未來的走勢作為投資的參考。但股市股票變化多端，難以以單一層面(技術層面)就能得出十分準確的預測結果。能影響一支股票的股價之因素可以分為：

獲利能力因素	純益率基本、獲利率、總資產報酬率
經營能力因素	總資產週轉率、固定資產週轉率
償債能力因素	流動比率、速動比率、利息保障倍數
資本結構因素	負債比率
公司治理因素	獨立董監事比率、董監事持股比率、大股東持股率
市場風險變數	BETA 值
其他因素	存貨週轉率、本益比、市價對帳面價值比

若能把以上因素正確加進決策樹模型當中，其預測準確率必定能大大提升。

4.3 參考文獻

[1] Quinlan, J. R.(1993), Programs for Machine Learning Morgan Kaufmann, San Mateo, CA.

[2] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., Classification and Regression Trees, New York: Chapman & Hall, 1984.

[3] Sorensen, Eric H., Keith L. Miller & Chee K. Ooi , "The Decision Tree Approach to Stock Selection,"Jounial of Portfolio Management, 2 7(1), 2000,pp. 42-52.

[4] Ou, J., and S. Penman. "Financial Statement Analysis and the Prediction of Stock Returns." Journal ofAccounting and Economics 11, 1989,pp. 295-329

[5] Levine, R. and S. Zervos "Stock Market development and long-run growth," The World Bank Economic Review, 10(2), 1996,pp. 323–339.

[6] Kagami, N., Iwamoto, R., and Tani, T. "Application of data mining method(ID3) to data analysis for ultra deep hydrodesulfurization of straight-run light gas oil- determination of effective factor of the feed properties to reaction rate of HDS." Fuel, 84(2-3), 2005,pp. 279-285.

[7] Heston, S. L., Rouwenhorst, K. G., & Wessels, R. E. "The role of beta and size in the cross-section of European stock returns." European Financial Management, 5, 1999,pp. 9-27.