The University of Hong Kong

Department of Computer Science

COMP 4801

Final Year Project 2017-2018:

Financial Data Forecaster

Interim Report

Date: 21/1/2018

Supervisor: Dr. C.L. Yip

Member: Ng Kwun Ting BEng(CompSC)

UID: 3035100897

Abstract

With the rapid development of computer technology, nowadays data scientists use lots of algorithms and scientific methods to do data analysis with computer. Within that, one of the famous topic is to predict the trend of financial market. Hence, Financial data forecaster is the core part of algorithmic trading in the finance industry. This project aims to build a program to predict the future value of financial products.

Acknowledgement

I would like to thank my supervisor Dr. C.L. Yip for helping and guiding me a lot throughout the process of my project.

Table of Contents

List of Figures

List of Tables

1. **Background and Objectives**

Background:

Investing on the financial markets is what a knowledge-demanded task that many professionals in the related fields and people are doing. With the computer technology nowadays, the term “Algorithmic Trading” is prevalent to the industry. In fact, there are lots of companies and individuals out there, who possess very complexing algorithms that can actually benefit from the financial market steadily. Although, there are many papers published to tell their trading algorithms, the really powerful ones are not published because of commercial secret and that is what we call the “Black Box”.

Objectives:

The goal of our project is to use scientific methods for building one of the powerful algorithms to forecast the financial data in the future. Hopefully, it can actually trade and benefit from the financial market, that is, “open the Black Box”.

2. **Theoretical Background**

The technologies used in this project involve the following theoretical background:

Pair trading:

Pair Trading is a method and trading strategy/concept that to find out a pair of stocks that are positive correlated. When there is a deviation between the two prices of the two stocks, there will be a force that the one (stock's price) gone above the average will go down and the other one(stock's price) gone below the average will go up until they reach an average/balance.

Artificial intelligence:

Artificial intelligence is a huge aspect that to design intelligent agent to complete a specific task optimally. In this project, the genetic algorithm, Simulated Annealing, artificial neural network(ANN) will be used as for optimization of the parameters.

Machine Learning:

Machine learning is a particular aspect of artificial intelligence, it mainly acts as a classifier to sort different items by types. In this project, classifier such as support vector machine(SVM) will be used to sort data to different types.

Time-series Analysis:

Time-series analysis is a hot aspect in statistics that considers the time-series data, based on regression methods. In this project, mainly two models will be used: Cointegration model - which proves the time-series data are positive correlated and "white noise" test - which has null hypothesis that the time-series data is a random process.

Technical indicators:

Technical indicators are developed by the finance industry to help understanding and predicting the trend of stock markets. It can be divided into two categories: trend indicators such as MACD and momentum indicators such as RSI and KD.

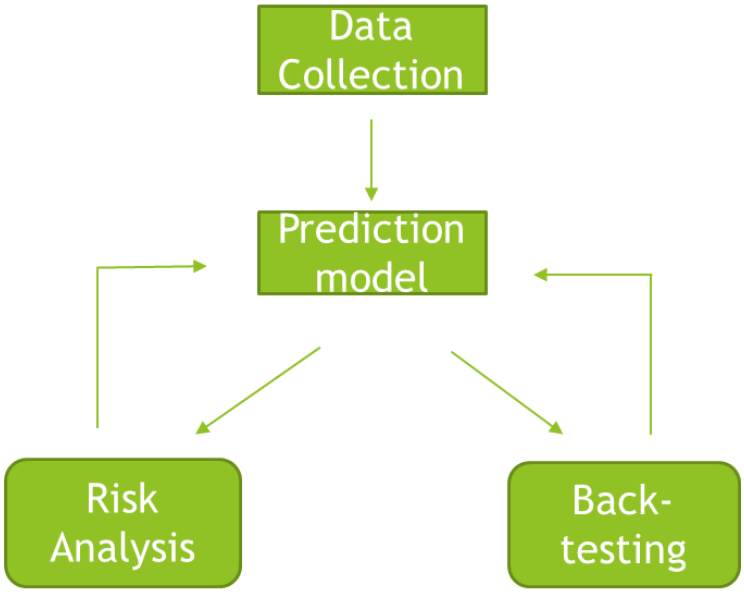
3. **Methodology**

3.1 General model concept

The general model of a financial forecaster is as follows:

Figure 1. General model

1. Data Collection
2. Prediction model
3. Back-testing
4. Risk Analysis



As by the efficient market hypothesis (Malkiel, April 2003) of some economists, the financial market has reflected all the information of factors on the price , and therefore , it cannot be predicted. However, some research (Khaidem, L., Saha, S., & Roy Dey, S. (2016)) state that the efficient market is a concept of the whole financial markets in different countries. Within the efficient market, there are a mixture of smaller efficient and inefficient markets and this is why many predictions come.

As I believe in the second claims above that there are small inefficient financial markets contained in the big general efficient financial market, I plan to develop below method to find out the predictable stocks, which integrates pair trading, technical indicators, machine learning, statistics and artificial intelligence methods.

3.2 **Description of the approach and logic of the designed method:**

The steps are as follows:

1. Because of the efficient market hypothesis(EMH), it is difficult to predict the market trend, therefore, I use the following algorithm, trying to predict the market.

2. First, I use the concept of pair trading to match some stocks in pairs by a time-series analysis model : Cointegration model , which can prove the two stocks are positive correlated statistically.

3. To this step, I know that, the pairs I find will have one stock's price(in a pair) go down and the other one's price(in a pair) go up in the future until the two prices attain an average(concept of pair trading).

4. Then I use 3 prediction rules as the trading signal, only if the data fulfil all the 3 criterions, the trading signal can then be generated.

5. The first prediction rule is the traditional strategy of pair trading, which uses the spread(the difference of prices of the pair) and the parameters from the regression in Cointegration model (alpha and beta) and the statistics from the spread of formation period (mean and standard deviation).

6. The second prediction rule is the most important part of my algorithm. It is mainly use the classifiers in machine learning (SVM, Adaboost) and use technical indicators(sush as RSI,KD,MACD) in the finance industry as the features to classify when is the entry point when the first prediction rule is fulfilled.

7. In order to further against the efficient market hypothesis(EMH), I use the "white noise" test (a "white noise" means a time-series is a random process that can not be predicted) to judge whether the trading period(time-series) is a white noise or not, if it is not a white noise, it means an autocorrelation existed for the stock tested in that trading period.

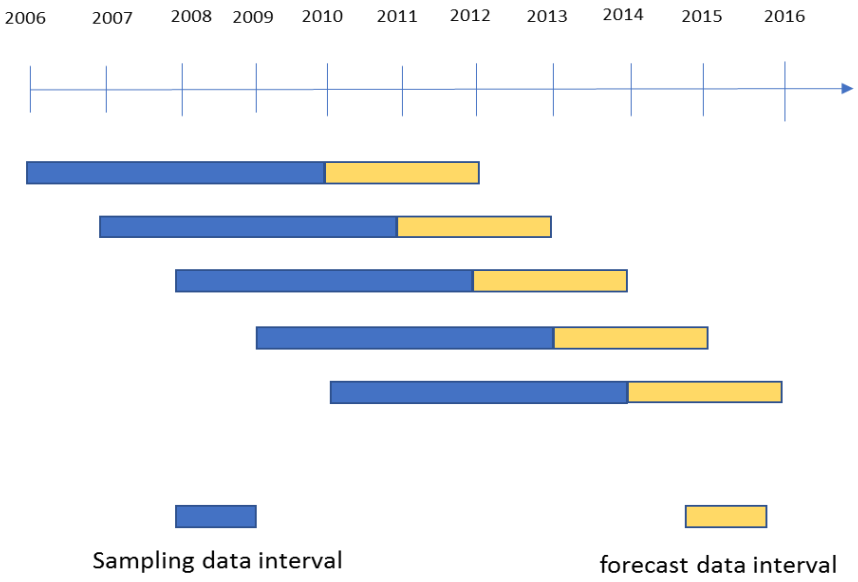
8. Optimizing the parameters in step5 (e.g. the traditional entry point 1.5sd) and step6(e.g. parameters and put/short signal of RSI) by AI algorithm.

9. To build a simple web app for the forecaster

3.3 Historical data testing method

The historical data is divided into two portions, one is for back-testing and fitting parameters, and the other is for the "Walk forward back-testing method" as the diagram below to reduce overfitting problem:

Figure 2. Walk forward back-testing method



4. Scope

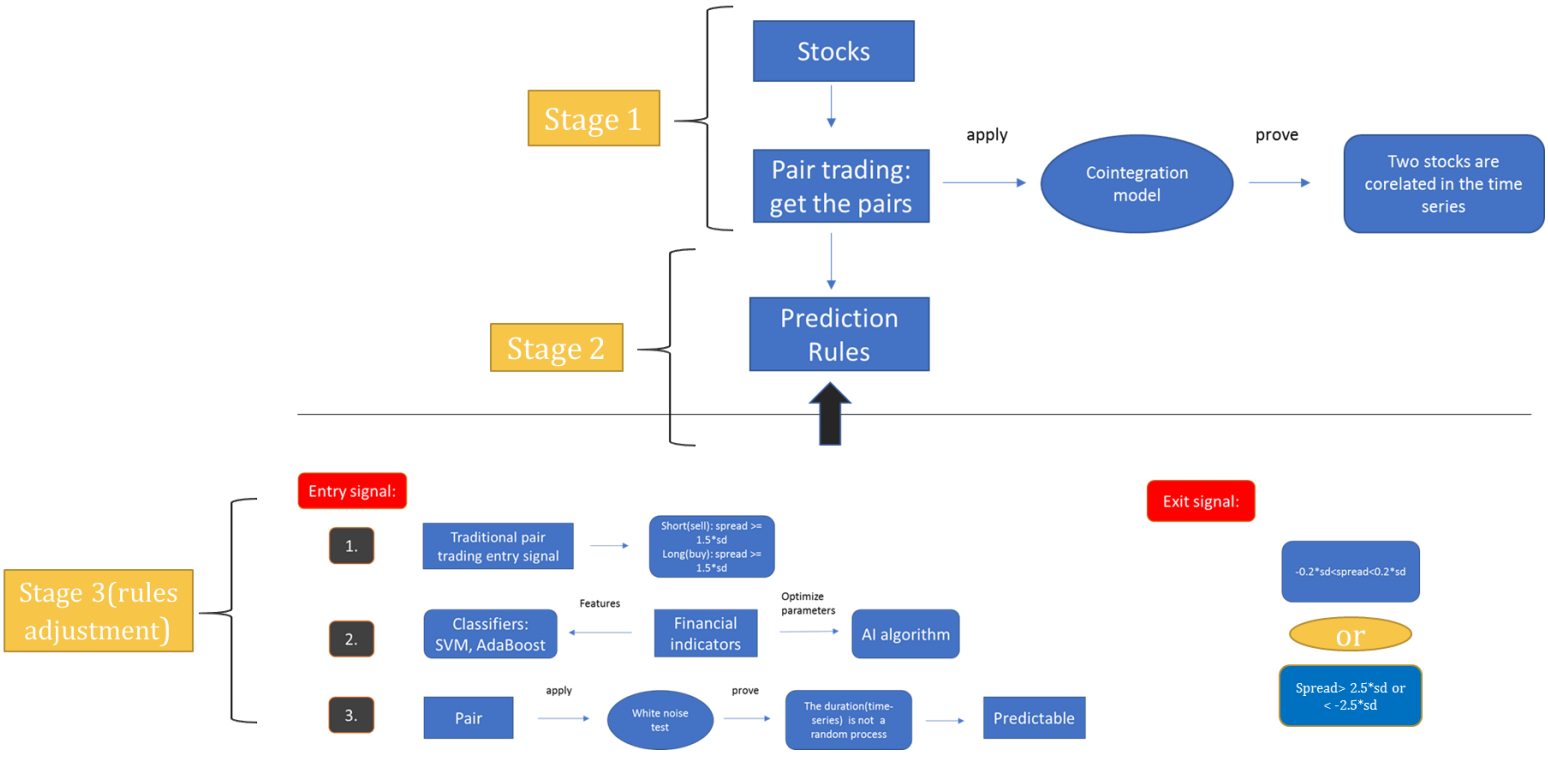
4.1 Data and language

The above is the main idea and flow of our forecaster, within the process, historical data (component stocks from HSI(Hang Seng Index), SSE50(Shanghai Stock Exchange Index), DJIA(Dow Jones Industrial Average) are used and collected from google/yahoo finance API. The forecaster is implemented by python.

4.2 prediction model architecture

In general, our prediction model scope (architecture) is as follows:

Figure 3. Our predictive model architecture



5. Deliverables

In total, there will be 9 deliverables:

1. Pairs of stock found from HSI,SSE50 and DJIA

2. main.py (main function to implement)

3. pairTrading.py (for defining a pair trading class which uses cointegration method)

4. tradingSignal.py (for generating trading signal)

5. backtesting.py (for simulating an account to backtest the result/profit/loss)

6. classfier.py (SVM and Adaboost)

7. techIndicators.py (contains around 15 technical indicators function for generating the indicator data as the features)

8. Result from the algorithm

9. A simple web app for the forecaster

6. **Challenges**

Optimization	-Each pair may have different optimized values for the parameters.
Fulfilling all the criterions	-To fulfil all the criterions we set, the number of final buying/selling signals may not enough to be proved as powerful for statistical significance
Assumptions may not be true	-Although the above model has some statistical tests in some steps such as those related to time-series to make conclusion, most of statistical models have some assumptions behind and therefore the conclusion in the corresponding step may not be true

Table 1. Challenges to face

7. **Project Schedule**

Steps:	Deliverables:	Date:
More researches on the technical details/information required in each step		Oct
Data collection / build web scraper for the classifier to reduce combinations tested	data needed, and (Web scraper)	Oct-Nov
Test the combinations and find out pairs of stocks, which fulfil the criterions, estimate the duration of the pairs	pairs of stocks found, which fulfil the criterions, and estimated durations of each pair of stocks will last for	Nov-Dec
Applying financial indicators to the specific duration, adjust/optimize the parameters	adjusted/optimized/mixed financial indicators	Dec-Jan
Find out more variables by web scraper/ factors calculated from financial industries, then applying PCA/FA analysis	(web scraper), and simplified (logistic) regression model	Jan-Feb
Compare and Combine the findings to get a more precise model	final combined algorithm	Mar-Apr

Table 2. Proposed project schedule

8. **Current Status**

The current status is I have finished the step 1-5 (Built the pair trading class for matching some pairs, first prediction rule and a simulation account for back testing), it is explained in the following content.

In the program, the pair trading uses the formation period as the formation of a pair and the trading period as the back testing.

The pair formation uses the cointegration model which has the core equation:

$$Spread_t = \log\left(P_t^Y\right) - \left[\hat{\alpha} + \hat{\beta}\log\left(P_t^X\right)\right]$$

Here, Y and X presents the two stocks in a pair and P is the price.

The value of alpha and beta can then be calculated by regression, which are then be used in calculating the spread of trading period.

Also, the mean and standard deviation of spread of formation period are calculated, which are used as the traditional entry signal (spread excesses mu +/- 1.5 sd) and exit signal (spread returns back to mu +/- 0.2 sd or spread excesses mu +/- 2.5 sd(the relationship of pair broken)) in the trading period.

Also, the program contains the Dickey–Fuller test (ADF) tests, which is used to test a time series data is weakly stationary(weakly stationary means there is autocorrelation of the time series itself and therefore it may be predictable). The ADF tests are used as a prerequisite for the two stocks to test before the cointegration model and there is also an ADF test in the cointegration model to test the spread.

After that, I have built a simulation account for backtesting, and the result is: there is some profits in general, but the amount of profit is very various and depends on the parameter values that I input.

For example, for a pair from SSE50 (no.: 601988, 600000) using the same formation period(i.e. same alpha and beta to calculate the spread of trading period) :

Trading period: 2015-01-05 to 2015-12-31

Entry signal: spread excesses mu +/- 1.5 sd

Exit signal: spread returns back to mu +/- 0.2 sd or spread excesses mu +/- 2.5 sd

```
=================== Backtesting in a trading period ====================

Trddt
2015-01-05    100000.0
Name: Asset, dtype: float64
Trddt
2015-12-31    143849.353626
Name: Asset, dtype: float64

Backtesting result:
cash used for put/short of each position: 20000
Asset at the beginning of the trading period: 100000.0
Asset at the end of the trading period: 143849.35362571833
Asset percentage change: 143.85 %
Profit gain: 43.85 %
========================================================================
```

Trading period: 2015-01-05 to 2015-06-30

Entry signal: spread excesses mu +/- 1.6 sd

Exit signal: spread returns back to mu +/- 0.3 sd or spread excesses mu +/- 2.8 sd

```
=================== Backtesting in a trading period =====================

Trddt
2015-01-05    100000.0
Name: Asset, dtype: float64
Trddt
2015-06-30    181264.543659
Name: Asset, dtype: float64

Backtesting result:
cash used for put/short of each position: 20000
Asset at the beginning of the trading period: 100000.0
Asset at the end of the trading period: 181264.5436590488
Asset percentage change: 181.26 %
Profit gain: 81.26 %
========================================================================
```

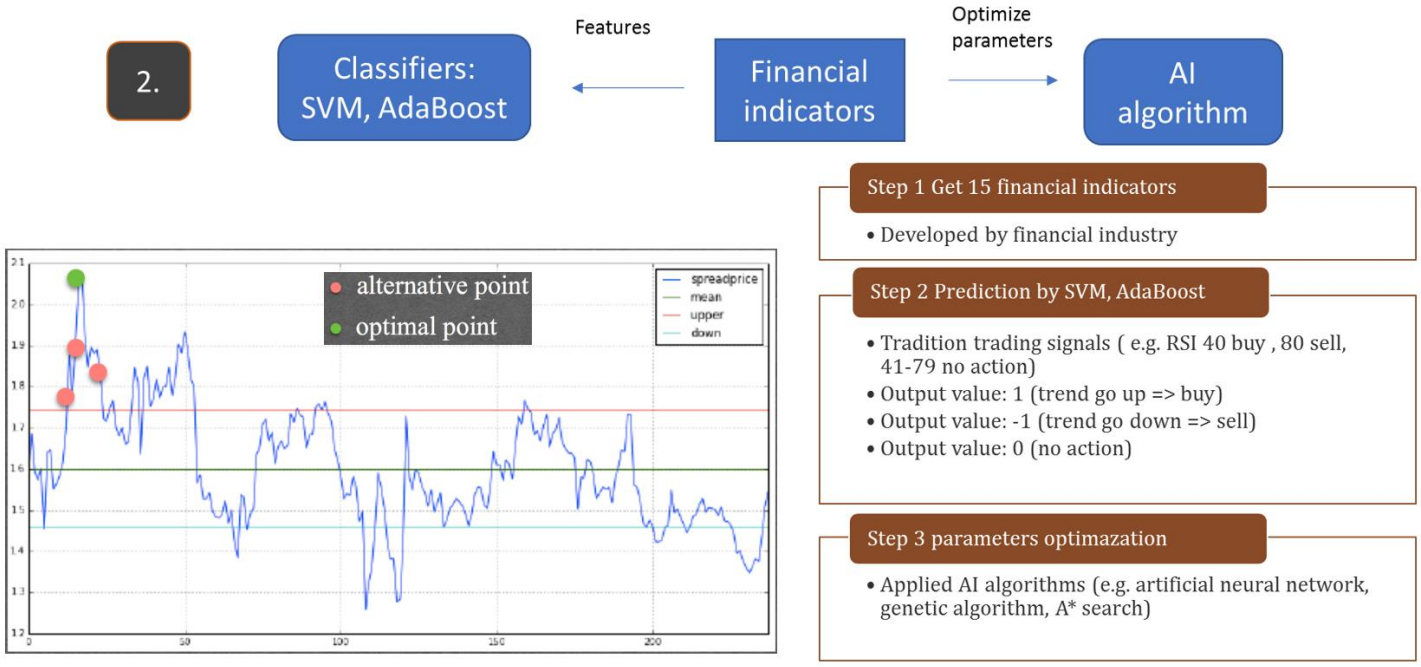
From the above result, we can see that by adjusting the parameters to different values, the trading period is halved but the profit is nearly double. (43.85% to 81.26%)

It means that only by the first prediction rule, the return is very unstable and mainly depends on the parameter values and trading period and some time may suffer small amount of loss. (Of course the return is for the assumption that we can put/short the stocks with the prices on the series and no extra cost is needed)

9. **Future work and summary**

The future worker will be the step 6-9 in the methodology.

For some explanations of step6 and 9:



By traditional pair trading strategy, the entry point is the red line (when the spread excesses mu +/- 1.5 sd), but the spread may continue to become larger in value and the optimal entry point is the peak (green point of the above graph). Of course, we don't know when is the optimal entry point, but by adding the second prediction rule (classifers, technical indicators as the features), we can obtain an entry point(red point of the above graph) at least as good as the traditional entry point because the algorithm must fulfil the above two prediction rules to generate the entry signal.

Summary:

Financial data prediction is a very attractive and challenging task that many people want to solve for years but is still a mystery(there is no sure winning strategy, at least, published for people). We know that to predict financial data in the future is an extremely complicated task and has a long way to go deeper and deeper. As ones of the people who want to solve this problem, we hope that through this project, we can build a possibly workable algorithm and find potential ways to go in the future researches.

10. Abbreviations

1. SVM: Support Vector Machine
2. AI: Artificial Intelligence
3. RSI: Relative Strength Index
4. KD: Stochastic Oscillator
5. MACD: Moving Average Convergence/Divergence
6. ANN: Artificial Neural Network
7. PCA: Principal Component Analysis
8. FA: Factor Analysis
9. P/E ratio: Price-Earnings Ratio

11. References

Malkiel, B. G. (April 2003). The Efficient Market Hypothesis and Its Critics. April 2003: CEPS Working Paper No. 91.

Introduction to Pair Trading -Based on Cointegration https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/inst/doc/IntroductionToPairTrading.pdf?revision=6&root=pairtrading&pathrev=6

The Application of SVM to Algorithmic Trading http://cs229.stanford.edu/proj2008/Blokker-TheApplicationofSVMtoAlgorithmicTrading.pdf