MINING TIME SERIES DATA FOR FINANCE APPLICATIONS

Project Proposal

Written by TONG Man Kin (08502621D)
Supervised by Prof. CHAN Chun Chung Keith, The Hong Kong Polytechnic University

October 7, 2010

Mining Time Series Data for Finance Applications

-

Project Proposal

Written by TONG Man Kin (08502621d),
Supervised by Prof. CHAN Chun Chung Keith,
The Hong Kong Polytechnic University

October 5, 2010

Contents

1 Abstract

The financial tsunami in 2008 has turned the world into a turmoil. Its destructive effect has lasted until today, causing people to reevaluate the risk of their derivative investments, and reconsider their current investment strategy. The regret of investors will certainly lead them back the basic strategy of investment, the old wisdom that we have forgotten. It is the part of a wise man to keep himself today for tomorrow, and not venture all his eggs in one basket[1]. It is called portfolio diversification.

The author believes a sophisticated portfolio optimization tool would satisfy the need of investors nowadays and help the promotion of rational investing. This tool is therefore proposed to be built. By applying data mining techniques, assisted by theories of financial mathematics, this scientific tool is considered able to suggest both the optimal stock selection and their optimal weighting in portfolios, and thus the investors are able to diversify the risk of their investments.

This project is expected to be completed in April, 2011 with estimated time 6 man-month. It will be done by the author of this proposal, and no special hardware or software resource is required from the department.

2 Background and Problem

This project aims at designing and developing a technique that is capable of discovering patterns in time series data. It also emphasizes on applications to ensure that the techniques developed can be used in finance for stock data analysis. After reviewing existing analysis techniques of stock data and understanding the investment behavior of individual investors, the author believes there is a lack of sophisticated portfolio optimization tools from them. Therefore, implementing a software tool which makes use of time series data mining techniques, and provides suggestions to users such that portfolios with better diversification could be built, is proposed for this project.

With a open and well-developed equity market, Hong Kong is flooded with a wide spectrum of derivative products, including different types of Equity-Linked Instruments (ELI), warrants, Callable Bull or Bear Contracts (CBBCs) and accumulators. Although accumulators are mainly traded by enterprises, other products have already been very common among individual investors. According to HKEx [1, 2], the percentage of warrants and CBBCs investors among adult population in Hong Kong has been greatly grown from 1% in 2000, to more than 12% in 2009. These products are often considered as useful instruments to hedge the risk of ordinary stocks, but seldom realized by individuals that they actually carry an even more complicated risk structure. Investors may suffer from total loss of capital by not knowing the characteristics of those derivatives, which is agonizing.

[1] Well-known dialogue from the novel "The Ingenious Hidalgo Don Quixote of la Mancha" by Miguel de Cervantes.

In the year 2008, the financial tsunami severely hit United States' economy, and the bankruptcy of a hundred-year financial institution, Lehman Brothers, spread its disastrous shock-wave around the world. Different kinds of derivatives issued by Lehman Brothers, turned out to be scrap papers in one night, along with the collapse of the company. Hong Kong, as one of the major market of these products, her investors lose all the money invested when they thought their investment are capital protected as the financial advisors assured [3]. Moreover, Lehman Brothers did not sell their derivatives directly to individual investors, instead it offered these products to those investors through various local banks of Hong Kong, like Bank of China and Hang Seng Bank, so most of these investors do not even know that the asset they bought is actually guaranteed by Lehman Brothers. Until now, 'Alliance of Lehman Brothers Victims', an organization formed by those investors, is still struggling through various means, including lawsuits and protests, trying to get a refund of capital from those local resale banks [3, 4]. By this tragic case, we learnt that ordinary investors do not have any idea on the characteristics of complex products which they are putting money in, even the actual issuer of those products. This comes into questions that, are derivatives products appropriate for individual invertors? If not, what kind of investment strategy should be applied as an individual investor?

Through buying a stock and its derivatives with bearish stance, or the opposite way, short selling a stock and buying its derivatives with bullish stance, earnings from those products would be able to cover some of the losses when the movement of that particular stock is out of the investor's expectation. Equity derivatives, therefore, are described as being able to 'hedge' the risk of investing the stock, that we could avoid excessive loss, in exchange, by sacrificing part of the possible return.

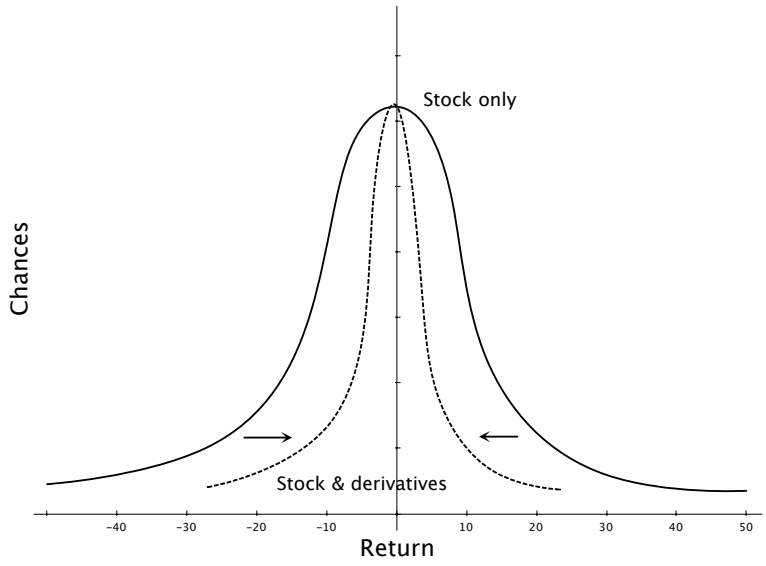

Figure 1: Illustrated return distribution of buying only the stock and buying both the stock and its derivative.

But this is not the end of the story. Although many of these derivatives are traded in exchange like ordinary stocks, they are guaranteed by third-party issuers. It may not be often, but there are still chances that the derivative issuers go bankrupt or break the contacts like the case of Lehman Brothers. Through 'hedging', market risk is reduced, but counter-party risk is created. Also, together with the complexity of derivative terms, like implied volatility[2] and barrier prices[3], making these products not suitable for normal investors.

In the author's opinion, instead of buying derivative products, the technique of portfolio diversification should be used to reduce the risk of buying stocks as an individual investor. There are mainly 2 types of risk in an operating business, they include systematic risk and business risk. Systematic risk refers to the factors common to all securities, like economic downturn, and business risk refers to the factors associated with individual assets, like projects that the business is doing, which is diversifiable. As a result, by selecting stocks with different areas, industries, business cycles and market trends to form a portfolio, price fluctuations of a single stock only contributes a low proportion of it, and thus the return of it will become more stable. Selecting a good portfolio can achieve the same advantage of trading derivatives, as shown in figure 1, without suffering from extra counter-party risk or studying complex terms.

There are existing theories and tools for optimizing the proportion of each selected stock in one portfolio, in order to achieve the maximum return per unit of risk among all other proportion options. However, it is still difficult to decide which stocks should be included in a portfolio. A sophisticated technique which is able to provide this kind of analytics and assist in users' investment decision making, as well as showing the importance of portfolio diversification to the general public, would be very beneficial to both individual investors and the society. By these backgrounds and reasons, the author is proposing this project to research on the application of data mining, to design and develop such a technique and software application.

3 Objectives and Outcome

3.1 Project Objectives

This project is targeted to research on existing data mining algorithms, design and implement a sophisticated software tool which can assist its users when they make investment decision. Users would enjoy the benefit of risk aversion when they study on, and then apply the stock combination advised by the tool. The study on data mining techniques will also benefit other software developers when they extend the capabilities of this tool or apply the same technique to achieve related goals. Last but not least, as one of the portfolio optimization tools available in the future, it intends to contribute in promoting the practice of diversification and rational investing to the general public.

[2]Often seen in warrants.
[3]Often seen in CBBCs and other barrier options like accmulators.

3.2 Project Scope Description

To accomplish the objectives stated in section 3.1, the following project specifications, scope and boundary is suggested:

- Software specifications
  - Contains an executable command-line component for End-Of-Day (EOD) stock data collection from Yahoo! Finance Hong Kong will be built. This component will accept input to specify the time frame of data to be collected, and save the data in user desired location.
  - Contains an executable command-line component which analyze the stock data collected by the above process will be built. This component will have several data mining algorithm built-in and allow users with relevant knowledge to select their preferred algorithm.
  - Contains a Graphical User Interface (GUI) for user interaction will be built. This component will visualize the analysis result, suggest additional stocks for the user's portfolio and calculate the optimal contribution of each stock in it. Statistical calculation of risk carried by the portfolio, and return per risk, will also be illustrated to user.
- Project scope
  - All stocks listed in main board or growth enterprise market (GEM) in Hong Kong, will be analyzed by the software tool.
  - Price data of these stocks, since their listing, or since the earliest date that the author could acquire their EOD data, will be analyzed by the software tool.
- Project boundary
  - All assets which are not listed in Hong Kong are excluded from the scope.
  - All other products listed in Hong Kong, like warrants, ELIs, CBBCs, debt securities, units trusts and mutual funds, are excluded from the scope.
  - Real-time and intraday stock prices are excluded from the scope.
  - Price data which cannot be acquired through the supported data source stated above, are excluded from the scope.

3.3 Project Deliverables

This project will be able to present the following deliverables upon its completion:

- Software
- Source code

- User guide
- Design specification[4]
- Project schedule[4]
- Test plan[4]
- Test report[4]

4 Project Methodology

4.1 Literature Review

There are certain theories and techniques proposed before for stock analysis and portfolio optimization, in this section, the author is going to discuss certain techniques which would be applied in this project.

4.1.1 Modern Portfolio Theory

Developed by Markowitz [5, 6], Modern Portfolio Theorey (MPT) is the first and also the most important foundation of mathematical techniques in portfolio optimization. It introduces the analysis of investment portfolios by considering the expected return and underlying risk of each individual assets and, crucially, the interrelationship of these assets in the portfolio. It provides a mathematical framework for quantifying risk and return, enables comparison between portfolios using these quantitative measurements. Before this, investors can only examine their investments one by one through fundamental or technical analysis, and then build up portfolios of their favored stocks without the concern of their relationship in between. Markowitz's contribution is a breakthrough in both the mathematical and finance areas at that time.

In general, risk and return are positively proportional. This means investments which have higher risk are expected to have a higher return. For example, investing on the real estate market would have a higher return than stock market, as the relatively lower market liquidity creates extra uncertainty, or risk, which requires a higher return to compensate. As a result, if there are two portfolios that offer the same expected return, there is no reason for investors to select the more risky one. With the belief on this condition that investors are risk averse, MPT calculates and compares the risk per return of given portfolios, so that investors can review these statistics and select a better portfolio.

The theory considers a portfolio as a weighted combination of its assets, and thus the return of a portfolio is the weighted return of its assets. Therefore, the expected return of portfolio can be expressed as the following formula:

$$E(R_p) = \sum_i w_i E(R_i)$$

[4]Design specification, final project schedule, test plan and test results will be included in the final report and submitted collectively.

R_p represents the return of the portfolio, R_i represents the return of individual stock, or investment i, w_i represents the the proportion of asset i contributed to the portfolio.

The theory also models the return of an investment as normally distributed and the risk of it as the standard deviation of return, and the variance of portfolio return is be expressed as:

$$\sigma_p^2 = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} w_i w_j \sigma_i \sigma_j \rho_{ij}$$

The second part of the formula is considered as a model of the interrelationship of two stocks and how this relationship affects the return of the portfolio, where ρ_{ij} is the correlation efficient between the return of the two stocks, i and j. And the standard deviation, or the risk of the portfolio, can be simply calculated by taking a square root of the variance:

$$\sigma_p = \sqrt{\sigma_p^2}$$

If an investor is provided with various portfolios for him to choose, he or she can apply the formulae above to calculate the return and risk of each portfolio and plot the following graph:

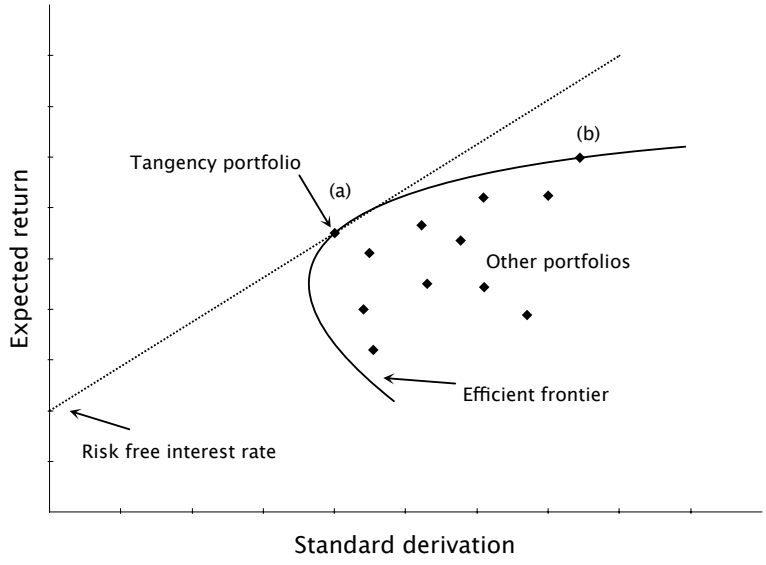

Figure 2: Illustrated efficient frontier. Portfolio (a) has the greatest return per risk among others, portfolio (b) has the same stocks as (a), but differs in each stock's weighting.

The curve which intercept with the straight line started from risk free interest rate, is called the efficient frontier. It is a portfolio which can offer the greatest return per risk, at point (a) of the graph, when certain weighting is applied on

its stocks. That particular portfolio and weighting of stocks is the best option available for the investor. He or she can simply reduce the holdings of this optimal portfolio and purchase risk free assets if lower risk desired, or simply borrow risk free assets to increase the holdings of this portfolio if higher risk and return is desired.

Instead of plotting out figure 2, the optimal portfolio can also be found by finding the maximum of return per risk of all available portfolios, the return per risk of a certain portfolio can also be calculated by the following formula:

$$S = \frac{E(R) - E(R_f)}{\sigma}$$

This is also called the Sharpe ratio. $E(R_f)$ indicates the expected return offered by the risk free asset.

MPT is considered effective to select an optimal portfolio among a few options, or deciding an optimal proportion for the given stocks inside the portfolio. However, it becomes very difficult to decide whether a stock should be included in a portfolio to achieve the maximum Sharpe ratio. Like Markowitz described [6], this theory is focus on the choice of portfolio, instead of the observation on the stocks.

4.1.2 Data Clusting Technique

Data mining is a technique from computer science, is the process of extracting patterns from data. It has becoming increasingly important and already been widely used in different business functions, like studying the customers' favorite, shopping behavior, or studying suspicious transactions to identify credit card fraud. Clustering, one of the streams of various data mining techniques, it focuses on measuring similarity and dissimilarity of data sets, and then grouping these data sets into different clusters. Data sets in the same cluster have similar data, where data sets in different clusters will have more distinct data.

One of the most common methods for measuring similarity and dissimilarity is euclidean distance. According to Gan, Ma and Wu [7], the formula of calculating euclidean distance is:

$$d_{euc}(x, y) = \left[\sum_{j=1}^{d} (x_j - y_j)^2 \right]^{\frac{1}{2}}$$

In the formula, x_j and y_j represents the j-th attribute of data set x and y, respectively. The greater the euclidean distance, the higher the dissimilarity between the two data sets.

After we know how the distance between data sets is defined, we need to know how we can group different data sets into clusters. One of the clustering algorithm, k-means clustering method, suggests that we could initially create those clusters by randomly selecting a data set for each of the clusters, and then for the rest of the data sets, we could compare the distance between them and the

mean of each clusters, and thus putting the data sets into the cluster where they have the shortest distance.

K-means and euclidean distance are the easiest and most popular way to perform clustering on data sets. However, due to their simplicity, they do not have a very high accuracy to process time-series data like stock prices. By this reason, one of the goal of this project is to study and find out a good data clustering algorithm for stock data processing.

4.2 Proposed Methodology

The author believes that the application of data clustering algorithms would be able to overcome the limitation of MPT. However, clustering technique may not be the best way to decide the optimal proportion of each asset in a portfolio which MPT is effective in. Therefore, both financial mathematics and data mining technique should be used in this project.

There are two steps to build a portfolio, the first step is stock selection. By various data clustering algorithms, we can study the fluctuation patterns of each stocks and group them into different clusters. By a combination of stocks from all clusters, the short-term fluctuations of each stock would be offset or mitigated by other stocks in the portfolio, and thus reducing the overall portfolio volatility. The second step is to decide the best proportion of each stock in the portfolio. Once the users of this software has selected their preferred stock based on our clustering result, we can use MPT to calculate the best weighting for them instantly.

4.3 Project Assumptions

By applying the above theories and data mining techniques, this project inherits the following assupmtions:

- There are patterns in the fluctuation of stock prices, stocks either benefited or damaged by the same market event, they will have a similar movement in stock price.
- Mean-variance statistical model, which consider return as normally distributed and express risk as the standard deviation of return, is effective to describe stock behavior.

5 Project Schedule

This project is expected to be completed in April, 2011 with estimated time 6 man-month. It will be done by the author of this proposal solely. The following table shows the estimated schedule of this project:

#	Start Date	Duration	Task	Prerequisite
A1	October 7, 2010	1 week	Final amendment on project specifications and approval	
A2	October 7, 2010	3 weeks	Further study on data mining algorithms	
A3	October 14, 2010	1 week	System design	A1
A4	October 21, 2010	1 week	Implementation of CMD data collection module	A3
A5	October 28, 2010	1 week	Implementation of CMD k-means clustering module	A4
A6	November 4, 2010	3 weeks	Implementation of CMD other clusting modules	A2, A4
A7	November 25, 2010	1 week	Design of GUI: data collection, clustering, view clustering result	A5 or A6
A8	December 2, 2010	1 week	Implementation of GUI: data collection, clustering	A7
A9	December 9, 2010	2 weeks	Implementation of GUI: charting library integration	
A10	December 23, 2010	1 week	Implementation of GUI: view clustering result	A9
A11	December 30, 2010	2 weeks	Preparation of mid-term progress report	

Table 1: Project schedule with estimated date and duration (term 1)

#	Start Date	Duration	Task	Prerequisite
B1	January 13, 2010	1 week	Implementation of GUI: view stock chart	A9
B2	January 20, 2010	2 weeks	Implementation of portfolio stocks weighting logic	
B3	February 3, 2010	1 week	Design of GUI: selecting portfolio	
B4	February 10, 2010	3 weeks	Implementation of GUI: selecting portfolio	A10, B1, B2, B3
B5	March 3, 2010	2 weeks	Program testing	A8, B4
B6	March 17, 2010	1 week	Project finialization	B5
B7	March 24, 2010	3 weeks	Preparation of final report	B6

Table 2: Project schedule with estimated date and duration (term 2)

6 Resources Estimation

As this project is focused on algorithm design and software development, no special hardware or software resource is required from the department. High speed computation equipments would facilitate the repetitive testing of algorithms but are optional. Below is a list of resources which are expected be used in this project, and are already available:

- Human resources
  - 6 man-month
- Hardware resources
  - 1 personal computer with network connection
- Software resources
  - LyX & LaTeX
    - Open source software for document processing
  - Oracle OpenOffice
    - Open source software for document, spreadsheet processing and graphing
  - Microsoft Visual Studio
    - Commercial software for software development
  - Microsoft .NET Framework
    - Commercial software for software development
  - StarUML

    * Open source software for system diagram illustration
- Omni Graph Sketcher
    * Commercial software for graphing
- Adobe Acrobat Reader
    * Free software for accessing electronic thesis and resource of library

References

[1] Hong Kong Exchanges and Clearing Limited, "Retail Investor Survey 2000," *Hong Kong Exchanges and Clearing Limited*, Mar 16, 2001. [Online]. Available: http://www.hkex.com.hk/eng/stat/research/ris/documents/oris00_e.pdf. [Accessed: Sep 22, 2010]

[2] Hong Kong Exchanges and Clearing Limited, "Retail Investor Survey 2009," *Hong Kong Exchanges and Clearing Limited*, Mar 30, 2010. [Online]. Available: http://www.hkex.com.hk/eng/stat/research/Documents/RIS2009.pdf. [Accessed: Sep 22, 2010]

[3] A. Lam, "RBS sued for sale of Lehman-linked derivatives," *South China Morning Post* (May. 13, 2010), sec. City2 col. City.

[4] Alliance of Lehman Brothers Victims, "Website of Alliance of Lehman Brothers Victims," *Alliance of Lehman Brothers Victims*, Sep 19, 2010. [Online]. Available: http://www.lbv.org.hk/. [Accessed: Sep 20, 2010]

[5] H. M. Makowitz, "The Early History of Portfolio Theorey: 1600-1980," in *Harry Markowitz: Selected Works*, H. M. Makowitz, Ed. New Jersey: World Scientific Publishing Company, pp. 5-16, 2009.

[6] H. M. Makowitz, "Portfolio Selection," *The Journal of Finance*, **7**, 1, pp. 77-91, Mar. 1952.

[7] G. Gan, C. Ma, J. Wu, *Data Clustering: Theorey, Algorithms, and Applications*. Philadelphia: Society for Industrial and Applied Mathematics, 2007, p.71 & 161.