Geometric Correlation Extraction Method for Intelligent Finance Data Analysis

XingXiu Ji

A thesis submitted to Auckland University of Technology
in fulfillment of the requirements
for the degree of Master of Computer and Information Sciences

June, 2012

School of Computing and Mathematical Sciences

Primary Supervisor: Dr. Russel Pears
Secondary Supervisor: Prof. Nikola Kasabov

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a University or other institution of higher learning, except where due acknowledgment is made in the acknowledgments.

Printed name: Xingxiu Ji

Signature:

Date:

Abstract

Trend forecasting could be one of the most challenging things in stock market analysis, as the data associated with stock is the time series data characterized in high intensity, noise, uncertainty, etc. In addition, beyond the time series data, from a macroscopic point of view, the influences of the external environment, such as national economy, technical progress, legal and political events, social and demographic factors, even the natural environment, are significant as well regarding the future trend forecasting.

The literatures show the successes of the machine learning methods in stock analysis, but at the same time, the literatures also show that, from the historical data of a stock observed simply extracting some rules that the stock follows, and thereby forecasting the stock movements within a certain future time period, is rather difficult and exhibits lower accuracy. However, while studying a stock with its graphical representation (zigzag curve), the researchers often found the similarities of the stock movements. Thus, a geometric method for extracting the movement similarity from the past data is developed this this research, the key improvement of the correlation extraction method proposed is the graphical trend similarity approximation, using an arc to represent the trend of a portion of the whole time series, the evaluation of the trend similarities is turned to calculate the distances of the two arcs of the two time series to the arc of the time series observed. This means, the new method extracts correlations from general variation trends, avoids extracting correlation form the time series directly, the mismatch problem is thus solved.

To verify the proposed method, based on SVR an experiment is conducted with the real world stock data. Based on the analysis results, the paper concludes that our method improves the performance of stock price prediction. Some future work, which is capable of promoting our method, has been discussed in the thesis.

Acknowledgement

I would like to thank all people who have helped and inspired me during my master study.

I especially want to thank my supervisors, Prof. Nikola Kasabov and Dr. Russel Pears, for their guidance during my research and study at Auckland University of Technology. Their perpetual energy and enthusiasm for research have motivated all their students, including me. In addition, Prof. Nikola Kasabov and Dr. Russel Pears were always accessible and willing to help his students with their research. As a result, research life became smooth and rewarding for me.

All my lab friends in AUT made it a convivial place to work. In particular, I would like to thank Lei Song and Yiming Peng for their friendship and help during my thesis.

My deepest gratitude goes to my family for their unflagging love and support throughout my life; this thesis would be simply impossible without them. I am indebted to my father, Guoqiang Ji, for his care and love. As a typical father, he worked industriously to support the family and spared no effort to provide the best possible environment for me to grow up and attend school. He has never complained in spite of all the hardships in his life. I cannot ask for more from my mother, Cairu Shan, as she is simply perfect. I have no suitable word that can fully describe her everlasting love for me. I remember her constant support when I encountered difficulties and I remember, most of all, her delicious dishes.

Last but not least, thanks to God for my life through all tests in the past years. You have made my life more bountiful. May your name be exalted, honored, and glorified.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

This chapter briefly introduces the background information, problem to be solved, proposed solution, motivation, and possible contributions of the research.

1.1The Problem to Be Solved

Since the first stock market in the modem sense was established in U.S almost two hundred years ago, people always try to extract the hidden relationships in the previous data so as to forecast the future trend. However, stock market is in a complex and non-linear dynamical environment, trend forecasting could be one of the most challenging things therein as the data associated with stock is the time series data characterized in high intensity, noise, uncertainty, etc. (Li, Hoi, and Gopalkrishnan, 2011), finding out the relationships in previous data sets to forecast the trend by means of computation has drawn a number of attention from researchers and specialists in the relevant industries.

Beyond the time series data, from a macroscopic point of view, the influences of the macro-environment, such as national economy, technical progress, legal and political events, social and demographic factors, even the natural environment, are significant as well regarding the future trend forecasting (Yung-Keun, K., Sung-Soon, C., & Byung-Ro, M., 2005). Therefore, the information contained in the time series data, reflecting the macroscopic factors beyond the stock market, must be recognized and considered in stock market analysis.

Thus, the problem to be solved by this research is about how to extract desired information, in order words the correlation knowledge, for trend prediction from the chaotic time series data of stock market in an appropriate way.

1.2 Brief Background

In the field of modern statistics or mathematics, in general the term “correlation extraction” is about finding the distance base similarities, or in more technical words, calculating the correlation coefficients, between two time series, via computational algorithms, such as Pearson’s correlation (Hong, 2009). However, while simply applying the standard definition of correlation to analyze time series data, e.g. the stock market data, the researchers found that the correlations, extracted for the stock movement trends based on every single time point, exhibit less usability. This is because in time series data, mismatches of the time points always occur (Jung-Hua, & Jia-Yann, 1996). For example, the stock market data is usually a zigzag curve in a graphical representation form, while comparing two curves

which have a similar trend, e.g. rapidly dropping, as they have a similar trend, the correlation coefficient calculated is expected to be high, however, as mismatches occur, the ordinary correlation calculation sometimes provides a relatively low result . That means the significant correlation knowledge may be ignored, thus the trend prediction associated with standard correlation extraction based on time series data is often deemed to be low in accuracy.

Resulting from above, new developments of correlation extraction mechanism based on its standard version to improve the usability of correlation extraction in time series prediction, in other words, to overcome the defect of which mismatches occur in correlation coefficient calculation, are necessary.

As such, correlation extraction in the field of time series dada prediction (more specifically, in this research, the stock market analysis), have been widely studied by the researchers and specialists. The importance of correlation extraction is thus widely recognized in the future trend forecasting in stock market. The computer-aided approaches for extracting correlation knowledge from the historical datasets of stock values are prevalently used since computers are widely spread over the world. Thus, the computerized programs based on mathematical algorithms to extract the correlations from the historical datasets and thereby to forecast the trends in stock market have become one of the most important measures for supporting the industrial experts' decisions in stock market investment (Zhang, 2009).

However, no stock market analysis program can guarantee a 100% accurate prediction. This is because the correlation knowledge extracted from the time series data of the stock market is not sufficient or significant enough. From a more technical point of view, the underlying computational (mathematical) methods used for correlation extraction is not sophisticated enough to support the highly accurate stock market trend prediction (Han, 2007).

On the other hand, in general, there are two primary types of approaches used for the stock market analysis (Chang, 2007): technical analysis approaches and fundamental analysis approaches, which both analyze the behavior of the stock datasets but from different aspects, so as to predict the trends in stock market for a variety of purposes. The effectiveness of these two types of analysis methods, which concentrate on different fields respectively, has been verified by practices, wherein technical analysis is applicable in

stock analysis for a shorter time trend prediction, and fundamental analysis is applicable in stock analysis for a long term trend prediction.

From above, it seems a combination of these two types of methods may be a perfect solution for both long term and short term predictions. However, the successful cases of using these two types of stock analysis methods in combination are seldom seen in practice.

1.3 Motivations

Previous research works regarding stock market analysis have shown the convincing evidences for which the correlation knowledge extracted from the chaotic time series data is helpful for machine learning based trend forecasting. Therefore, from a technical point of view, methods for correlation extraction from time series data, or in order words, how to extract useful correlation information/knowledge from existing time data for trend prediction, is a crucial factor for accurate prediction, and thus has become a research hotspot nowadays.

However, while acknowledging the importance of correlation extraction in trend prediction, the previous research works regarding stock market analysis also show the insufficiencies of the existing methods of correlation extraction in stock market trend prediction (Han, 2007). In detail, the existing correlation extraction methods may be unable to provide satisfactory prediction results in terms of accuracy, or unable to provide versatility for use in different sectors, for example, for both short term and long term prediction, due to its inherent drawbacks.

A scan on the literature shows a gap between the theory and practice exists, as despite the usability and applicability of correlation knowledge in stock market has been accepted, there are still very few studies conducted on real world data for stock market analysis, with regard to the correlation extraction aided machine learning prediction.

Resulting from above, from a unique perspective, e.g. a geometric perspective, providing an improved correlation extraction approach and verifying the method provided with the industry-wide prevalent machine learning algorithms, e.g. SVR, is a fruitful area for research. More importantly, such a research could provide precious information for those who are committed to improve the accuracy of trend prediction in stock market analysis.

1.4 Research Objectives and Contributions

A primary goal of this research work is to work out a novel correlation extraction method for extracting correlation knowledge important for accurate trend prediction in stock market, the correlation extraction method is expected to have the advantages form both technical analysis and fundamental analysis, i.e., the new method should be able to consider the influencing factors from macroscopic environment in addition to the historical data itself.

A sophisticated literature review is thus necessary in prior to the development of the method, for helping the researcher have a comprehensive understanding to the status quo of the field of stock market analysis, and thereby develop a method which meets the expectations set for the research.

Another goal of this research is to examine the correlation extraction method with real world stock market data by means of a prevalent machine learning algorithm. The prediction using the proposed correlation extraction approach in conjunction with the selected mathematical algorithm is expected to provide more accurate prediction results, as compared to the application of only using the mathematical algorithm itself.

Therefore, the contributions of this research may include:

- Developing a novel correlation extraction method, this method takes into account both the microscopic and macroscopic factors while extracting correlation knowledge; and
- Verifying the correlations gained by the correlation extraction proposed by using machine learning prediction based on real world stock data.

1.5 Brief Description of Solution

From the historical data of a stock observed simply extracting some rules that the stock follows, and thereby forecasting the stock movements within a certain future time period, is rather difficult and exhibits lower accuracy. However, while studying a stock with its graphical representation (zigzag curve), the researcher found that the similarities of the stock movements often exist. Thus, a geometric method is developed to extract the movement similarity from the past data. With this geometric approach, an arc is modeled to approximate graphically the zigzag curve of a specific time length e.g. 20 days (in other words it is about using a continuous function is used to represent discrete data approximately). Thus there are in total n (n1-n20, n2-n21…) arcs drawn within the whole time range of the historical data (while an arc, as a part of a circle, is defined, the center

and radius of the circle are determined therewith), and then SVR (Support Vector Machine Regression) is used to predict the movement of the stock (market trends) based on the correlation knowledge represented by these circles.

The time series within the specific time length is the dataset used to generate rules in SVR, the similarity between two time series is determined by calculating and comparing the respective distances of these two time series to the observed time series (i.e. simply speaking, to find out the trend of the most similar time series).

The key technology of the correlation extraction method proposed is the graphical trend similarity approximation, using an arc to represent the movement of a portion of the whole time series; it is called geometric correlation extraction method in this research. The trend similarity between two time series (two portions of the zigzag curve of a stock), is modeled as the trend similarity between two arcs graphically approximating the two time series, this method is taken as a core factor for extracting correlations.

A weighted Pearson's correlation algorithm could be used to measure the strength of the correlations (Hui, & et al., 2004), thus the correlations with other factors, such as with correlative stocks or macroscopic economic data, are introduced to assist the trend prediction, and have different weights. Therefore, it is possible for the correlation method proposed to combine the technical analysis and fundamental analysis, concerning both microscopic and macroscopic factors, for enhanced market trend forecast.

The prediction implemented in this research is based on the proposed geometric method and SVR. The data required in this research project is stock prices over time and collected from electronic resources. The proposed method will be examined by comparing its results with the results of SVR on its own, thus, several accuracy metrics are used to measure the results.

1.6 Structure of the Thesis

To present the entire research in a systematic and sophisticated way, the research report consists of 6 chapters in total, wherein:

Chapter 2 is a literature review giving an overall impression of the status quo of stock market analysis, by reviewing the literatures in relevant fields. In more detail, within this chapter, the traditional approaches for trend prediction in stock market, including technical analysis methods and fundamental analysis methods, are introduced and discussed.

Chapter 3 contains a brief introduction of the machine learning algorithm for trend prediction, SVR, which is used in this research as the computational basis. In addition, this chapter also briefly describes the accuracy metrics which will be used for measuring and comparing the results.

Chapter 4 describes the methodology for conducting this research, within this chapter the proposed correlation extraction method and the data collecting and analyzing process are described in detail.

Chapter 5 presents the experiment results; the researcher compares the results of the SVR in conjunction with the proposed correlation extraction method with the results of SVR on its own using the accuracy metrics, and discusses the difference.

Chapter 6 concludes of the research and gives the directions for future work.

Chapter 2: Literature Review

This chapter mainly reviews the technical analysis, fundamental analysis methods, and other machine learning methods used in finance maket analysis.

2.1 Stock Market Analysis

In general, in stock market there exist two primary types of approches used for trend prediction: technical analysis approches and fundamental analysis approaches , both of which analyze the behavior of the time series data, i.e, the stock prices varied over time, but from different aspects, to predict the stock market trends/movements.

2.1.1 Technical Analysis

In brief, the technical analysis approaches are based on some particular mathematical techniques to predict future trends through studying the past market data. Technical analysis commonly presents predicitions or advices for trading in the form of chart, but usually does not consider the macroscopic environment (Nison, 1991). A number of technical analysis related approaches, such as running averages, candlestick charting, Dow theory, Elliott wave theory, Miscellaneous patterns, etc., (Neely, 1998). have been developed and widely used in stock market for a long period to provide trend prodictions for supporting people's trading decision, via a number of graphical reprentations (mainly charts).

For instance, as a prevalent approach for small investors, running average in stock market analysis could be defined a "finite impulse response filter" for analyzing time serie data (stock prices) through caculating the averages of the sub-datasets of the entire data series (Tabbane, & Mehaoua, 2004), in other words, dividing all the given time series data into different groups and calculating the avarage of each group, then connecting all the averages in line to respresent the market movements. Possible trends thus can be observed from the chart. A running avarage chart is suitable for forecasting the moving trend of the stock in a short term period. In addition, the particular subsets in the time series could be given specific weights for emphasizing special factors which may have particualr impacts on the stock movement (Neely, 1997).

For example again, candlestick charting is another prevalent tool for stock market trend forecasting, it is a type of graphical approache primarily using bar-charts to describe the stock marcket trends. A candlestick chart consists of a number of candles indicating the movement of a observed stock, eah candle is defined by the opening, highest, lowest and

closing prices of the observed stock in each of the trading days. Based on the theory described above, a weighted candlestick charting method called “Heikin Ashi candlesticks” is developed, and uselly used with the regular candlestick charts, to emphasizing some spealcial sub-datasets. (Nison, 1991)

In stock market analysis, Dow theory is another prevalent method for presenting stock price movement. It involves some definitions related to “sector rotation”, wherein sector represents a group of companies of a similar business. It is based on serveral princples: the market has three movements: main, medium swing and short swing movements; the market has three phases: accumulation, public participation and distribution phases; the market ignores all news ; the market have confirmed averages; the market has confirmed trends (Goetzmann, 1997).

2.1.2 Fundamental Analysis

Different from the technical analysis methods describe above which only focus on the stock price data per se, the fundamental analysis methods tend to examine the financial and operating conditions of the company owning the observed stock. What the fundamental analysis methods consider for providing stock trading advices may include a number of factors beyond the stock data per se, such as historical sales, net profits, assets, debts, industrial prospect, competitive intensity in the market, etc. (Dodd, 2009). However, it is noted that the factors taken by fundamental analysis method may include the microscopic factors which are directly related to the company per se, such as sales and profits, and macroscopic factors directly or indirectly related to the macroscopic environment, such as legal or political changes in national economies.

Since the fundamental analysis methods examine a variety of factors beyond the stock data per se, the factors used can be classified into two categories: quantitative and qualitative.

In stock market analysis, as described above, fundamental analysis takes both its value and the related qualitative or quantitative economic and financial factors into account to indicate the movements of the observed stock and thereby to provide help in decision making. Therefore, in simple words, the fundamental analysis methods could be defined as “researching the fundamentals” of the company owning the observed stock, the outcomes of a fundamental analysis may include the combination of graphical representations and texts (Dodd, 2009). Compared to the technical analysis methods, the fundamental analysis methods are able to provide extra insights which are extracted from macroscopic

environment and have impacts on the movements of the observed stock. For example, fundamental analysis is able to predict the trend of an observed stock in a time horizon by analyzing the strengths of the company owning the stock and the economic trends in the market in which the company competes, determining whether the market is going to grow or shrink, whether the current stock value is underrated or overrated, and finally synthetizing a sophisticated result by examining all possible factors.

In detail, some primary terms used in the fundamental analysis for defining a stock with respect to its future movement are listed below (Glass, 2008):

- Value stocks, for those of which values are underrated;
- Growth stocks, for those which are considered as having growing earnings;
- Income stocks, for those of which returns are steady;
- Momentum stocks, for those of which values may increase rapidly but are uncertain in a long term period.

Compared to technical analysis, as taking macroscopic factors into consideration, fundamental analysis is considered more reliable for use in forecasting a long term trend of a stock; however the results of the fundamental analysis method may not be as intuitional as the technical analysis methods.

2.2 Computational Finance

Nowadays, computers have been widely used to extract information from raw data for supporting human's decision making. Machine learning and prediction is such a type of automated approache allowing computer to study the behavior of the datasets so as to provide predictions.

2.2.1 Machine Learning and Prediction

Recently, machine learning technology which studies raw data by means of specified mathematical algorithms to extract information/rules that the data may follow, and thereby to make prediction has been widely used in stock market analysis.

A number of approaches and algorithms, such as Support Vector Regression (SVR), Linear Regression (LR), Neural Networks (NNs), Genetic Algorithms (GAs), and such like with their applications, have been widely used and their performances have been compared (Vapnik, 1998).

In the 1990's in many research studies, NNs were widely accepted as the best mathematical algorithm in use for making prediction based on time series data, which has

been proven by many research studies udndertaken. This is because the NNs methods are characterized by non-linear curve fitting (Hopfield, 1982), while most other algorithms may encounter over-fitting problems. That means, except for NNs, the use of other mathematical algorithms in non-linear curve fitting may lead to underestimated or overestimated values, deviating from the expected curve, as these algorithms in nature follow the extremes of the fitness functions rather than the average (Shin, Lee, & Lim, 2005; Yoo et al., 2005).

The over-fitting problem in Support Vector Machine (SVM) is not that serious in some way; and therefore has been another prevalent algorithm for trend prediction in stock market. However, a single SVM is unable to provide high accurate prediction results for stock market analysis (time series data). Literature show that the best prediction results for SVM are less than 73% in accuracy (W. Huang, et al., 2003). That forces researchers and specialists to seek for complementary solutions. Under such a circumstance, integrating SVM with other algorithms has been a new research direction. In literature, a lot of such studies are evident, combining SVM with other machines, such as back propagation neural networks (BPN), wrapper voting machines (WVM), etc. for higher accuracy is very common (W. Huang et al., 2004).

Support Vector Machine Regression Model is such a SVM based model which is developed by Tay and Cao (2002). This method was used to forecast the trends in the Australian Forex market by Kamruzzaman, Sarker, & Ahmad (2003). The experiment results show that SVR based computational method is able to effectively improve the prediction accuracy. (More about SVM/SVR will be elaborated in Chapter 4)

2.2.2 Correlation Extraction

Correlation represents the strength and direction of the relationship between two time series. In general, correlation can be grouped into two categories: linear correlation and non-parameter correlation. As a linear correlation extraction method, Pearson's correlation is one of the most well-known correlation extraction methods currently used for stock market analysis.

2.2.2.1 Linear Correlation

In Pearson's correlation, the Pearson product-moment correlation coefficient $\rho_{X,Y}$ is calculated via the following equation(Pearson, 1897):

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

wherein: X and Y are time series, wherein X = {x_1, x_2, . . . , x_N} and Y = {y_1, y_2, . . . , y_N}, cov (X, Y) is the covariance of X and Y, μ_X and μ_Y are expected values, E is the expected value operator, σ_X and σ_Y in the denominator are standard deviations. Pearson's correlation also provides a calculated probability p-value, which is associated with the Pearson product-moment correlation coefficient $\rho_{X,Y}$. The p-value can be calculated via the following equation:

$$\mathrm{p} = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathrm{p}_i$$

wherein,

$$\mathrm{p}_i = \begin{cases} 0 & \text{if } \Delta x_i > 0 \text{ and } \Delta y_i > 0 \\ 1 & \text{if } \Delta x_i < 0 \text{ and } \Delta y_i > 0 \\ 1 & \text{if } \Delta x_i > 0 \text{ and } \Delta y_i < 0 \end{cases}$$

In addition, as $\sigma_X^2 = E[(X - E(X))^2] = E(X^2) - E^2(X)$, $\mu_X = E(X)$, and $E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$. the equation for $\rho_{X,Y}$ can be revised into:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

wherein the equation is subjects to $\mathrm{p} < 0.05$.

From the equation, the calculated $\rho_{X,Y}$ is in the range of +1 to -1, that means Pearson's correlation could be positive or negative correlation, wherein positive correlation means that, one of the time series Y becomes large following the other time series X, and vice versa. While$\rho_{X,Y} = 1$, a perfect positive linear relationship between time series X and Y is reached. In contrast, while $\rho_{X,Y}$ is negative, while one tine series X increases the other time series Y decreases, and vice versa.

It should be noticed that Pearson's correlation $\rho_{X,Y}$ works only when the p-value is less than the significance level α, which typically ranges from 0.01 to 0.05. Pearson's correlation could provide higher accurate prediction results, in particular when correlation amongst time series is strong. The demonstration of use of Pearson's correlation in finance market analysis for forecasting trends is presented in the research work of Nicewander (1998). The

prediction result of Pearson's correlation in conjunction with NNs in their experiment, for predicting the trends in exchange rates between foreign currencies proves the applicability of Pearson's correlation in finance market analysis.

Another study conducted by Kwapien, et al. (2009), tests the performance of NNs in conjunction with Pearson's correlation, the results show that the combination of NNs and Pearson's correlation has higher accuracy for prediction based on exchange rates as compared to the use of NNs on its own.

However, the relevant literatures also show that Pearson's correlation is not as reliable as expected; it is significantly influenced by outliers, unequal variances, and non-normality in the underlying data distribution (Song, 2009).

2.2.2.2 Non- parametric Correlation

Different from linear correlation, another common type of correlation extraction methods, non-parametric correlation, is defined as correlation independent of parameters, for example, the statistics based on the ranks of observations, or applying Pearson's correlation to calculate the ranks of the data, instead of the data values per se.

There are a number of specific non-parametric correlation methods, such as Chi -square correlation, Spearman's correlation, etc. that are well-known in the industry, and may be used under the circumstances where the data only contains rankings but lacks numerical interpretations. (Song & Zhang, 2008)

2.2.3 Applications of Machine Learning in Stock Market

Heping Pan (2003) et al, conducted a research project on professional, technical and academic quantitative analysis of the finance and stock markets in 2003, aiming to bridge "the deep gulf between the two fields and unifying them under a general science of intelligent finance or financial intelligence". He found that recently the focus on the research of analysis and prediction in finance market were mainly placed on chart patterns and technical indicators, only involving some simplistic use of technical analysis and consequently leading to lack of precise and usable information for prediction. He also pointed out in his research paper that currently the technical analysis and fundamental analysis is converging. The final outcome of the convergence of the technical analysis and fundamental analysis could be a unified analysis approach of so-called "intelligent finance". This convergence exists because, as described before, the financial market in particular the stock market is in a complex and non-linear dynamical environment, the trading systems are built and supervised by people. In addition, due to globalization nowadays there is no

completely closed stock market in the world, the stock market is tightly bundled with the global economy and politics (as we discussed before, stock market is significantly influenced by the macroscopic factors). Resulting from above, Pan asserted that it is impossible for finance analysis to be completely objective, the empirical financial analysis is more practical. (Pan, 2003)

Based on the theory mentioned above, a market model called Swingtum theory is proposed (Pan, 2003) , which is based on an assumption that the stock market as a whole can be represented by a benchmark index, such as the NZX 50 index for New Zealand. The curve of a stock index consists of four types of fluctuations: dynamic swings, physical cycles, abrupt momentums and random walks, wherein the dynamic swings represent the long term business cycles, e.g., 3-5 years, it can be modeled by power laws; physical cycles represent the constant cycles, such as monthly cycles or weekly cycles, it can be modeled by Hilbert transform.

Another researcher, Sawaya (2005), conducted a research project aiming at “combining the Chaos theory postulates and Artificial Neural Networks classification and predictive capability”, presents his efforts in the field of financial time series prediction. With his method, Chaos theory is used to provide qualitative and quantitative tools for prediction, wherein quantitative tools provided by Chaos theory determine whether the time series observed is predictable; at the same time the qualitative tools provided by Chaos theory give further observations on the predictability of the observed time series. If the time series is considered as predictable, phase space can be reconstructed by time delays embedding according to Takens’ embedding theorem, thus multiple embedded vectors are generated. A cognitive approach is introduced into this research work for direction prediction for the time series observed. Therefore, the separation and embedding dimension in phase space reconstruction is determined by appropriate methods, such as False Nearest Neighbor. Then, the multiple embedded vectors obtained in a previous step are classified using appropriate methods; in this case Fuzzy-ART is suggested. Later, the algorithm BPNN (Back Propagation Neural Network) is trained with the embedded vectors for prediction.

Kwong (2007), with his colleagues proposed a method for financial prediction using an advanced paradigm from computational intelligence: Chaotic Oscillatory-based Neural Networks (CONN), in conjunction with the use of a fuzzy membership function. The method is based on the time series data in finance market data to predict market trends in a certain time horizon. The researcher stated that using a chaotic oscillator, such as the Lee

Oscillator, as the activation function of NNs is able to reach higher prediction accuracy than using a traditional hyperbolic tangent function in NNs. The forecasting model developed by Kwong and his colleagues for financial analysis is illustrated as the following figure (as in Figure.1):

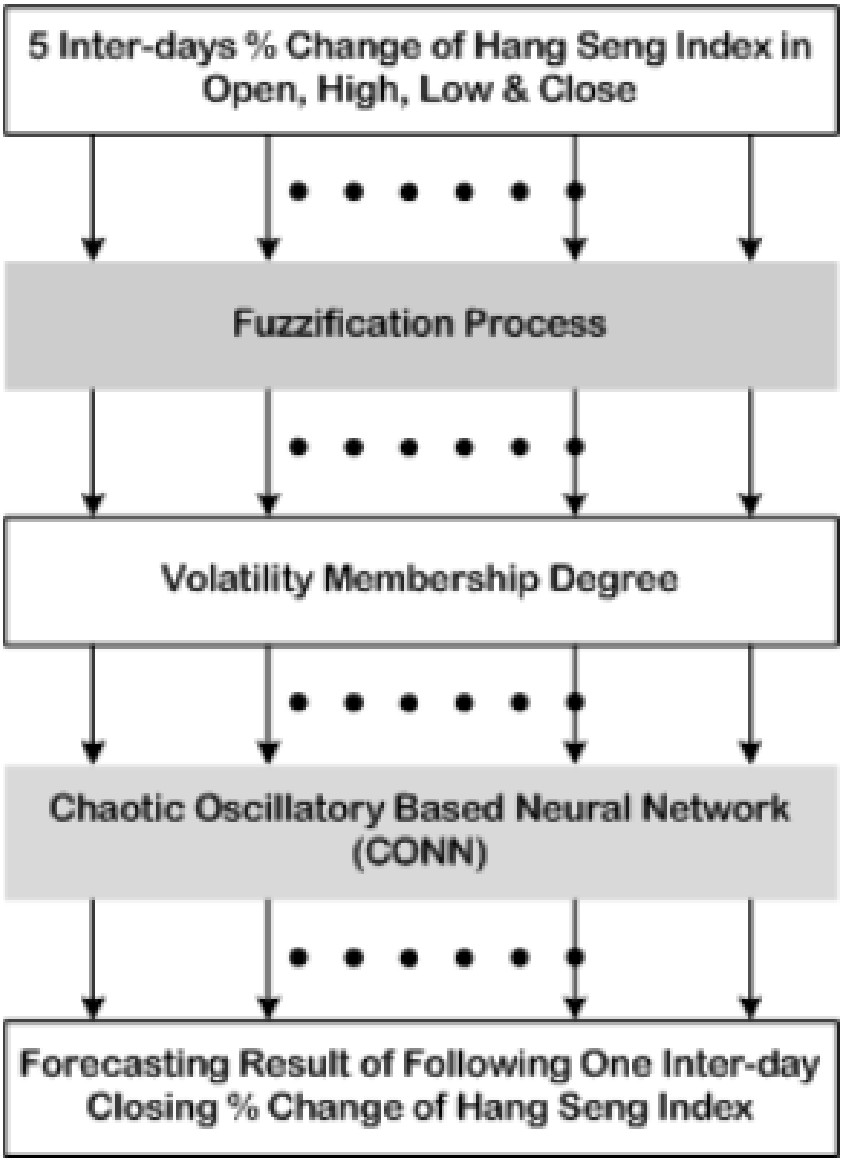


Figure 1 the forecasting model

The research compares prediction results of Chaotic Oscillatory-based Neural Networks (CONN) and the ordinary NNs for finance market analysis, the values of the Hong Kong HangSeng Index (HSI) between 1990 and 1998 are used as the training data for CONN and NNs, and the values of the HSI between 2007 and 2008 are used as the testing data in their experiment. The results of the experiment show that their model is able to reach higher accuracy in prediction the trends of the HIS, compared to the ordinary NNs.

Resulting from above, the prediction model proposed by Kwong and his colleagues is able to provide valuable information for supporting people's decision making, and thus suitable for the use in trend prediction based on time series.

In another piece of research (Abraham, et al., 2001), technical analysis is used to analyze stock markets; the authors set forth the application of their hybridized automated computing techniques for stock market analysis and trend prediction. The authors use NNs algorithm to forecast the stock value market for a short-term time period (one day for example) and develop a neuro-fuzzy system to further forecast the trend of the stock observed. The proposed technique is illustrated by the following figure:

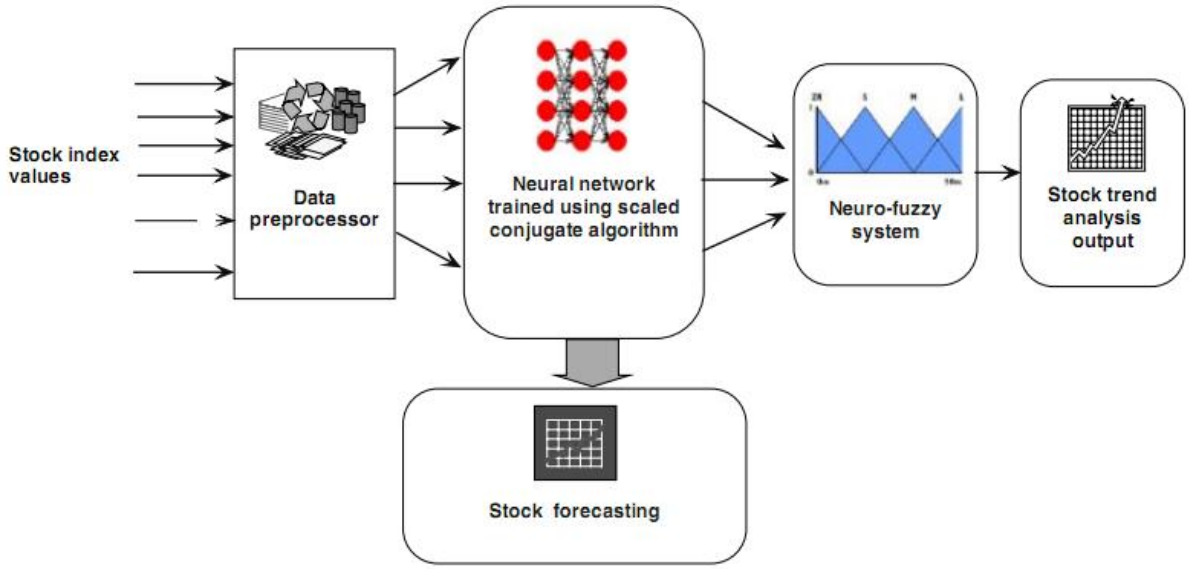


Figure 2 the proposed methodology

To examine the performance of the proposed technique, the researchers use the NASDAQ -100 index of the NASDAQ Stock Market. The data for NASDAQ -100 index of 24 months, and the stock data for six companies random selected from the NASDAQ -100 within the same period used as the training data are fed to the Artificial Neural Networks (ANNs) of the prediction technique proposed for prediction, after pre-processing by principal component analysis. Then, the stock values generated by ANNs as the prediction results are further analyzed in the neuro-fuzzy system for forecasting the trend of the market. In order to improve the accuracy of the prediction, the researchers also use ensemble ANNs rather than a single network.

In addition to the above, another prevalent algorithm, Self-Organizing Maps (SOMs), is also used by some researchers to forecast stock market trends, and its performance is tested in

many studies. SOMs algorithm is a type of unsupervised learning algorithms, with which the users do not need to know the answers prior to the prediction run, in contrast to ANNs, which belongs to supervised learning algorithms. Thus, SOMs is more suitable for the use in “finding or creating patterns that summarize and store useful aspects of our perceptions” in their research conclusions. (Bantouna, Tsagkaris, & Demestichas, 2010)

Chapter 3: Theoretical Foundations: Support Vector Machines

3.1 Introduction

As one of the most prevalent supervised data learning methods in the field of machine learning, Support Vector Machines (SVM) is widely used for trend prediction based on time series data, e.g., for forecasting in stock market. In the field, Support Vector Regression (SVR) is a variant of SVM for regression.

Just like other supervised learning algorithms, SVM analyzes and classifies the data input, however, it is noted that, SVM has its own characteristics. SVMs work on the principle of finding the maximum margin hyperplane between two classes. Such hyperplane can be either linear or nonlinear, depending on the type of kerned function employed. Maximum margin hyperplane are constructed with the distribution of support vectors which are represented by data instances in close proximity to the hyperplane. Thus the SVM model can be represented by a small set of support vectors and is thus less vulnerable to over fitting than other types of machine learning algorithms. SVM's have been widely applied in classifications, but recently Support Vector Regression (SVM) based on SVM has been developed for numeric prediction.

3.2 SVM and SVR

For a standard SVM, the training data input, or a set of samples for training in other words, are classified into two categories first; and then the classified samples were trained by the training algorithm in SVM to build up a model; the newly input samples are classified by the model into the categories, as that, SVM is considered as a non-probabilistic binary linear classifier. The model built by SVM training algorithm is used to assigns the new samples by mapping the examples as points in space, so as to divide the examples into two separate categories. Thus, the prediction for a new sample is about which category the new sample should belong to.

Initially, SVM was developed as a linear classification algorithm; in 1992, based on the effort of other researchers, SVM was extended into the field of non-linear classification, the researchers enabled SVM for non-linear classification via polynomial functions and Gaussian RBF, which are often known as kernel functions, instead of the linear hyper planes in the algorithm (William, et al., 2007).

In the use of SVR for prediction based on time series, a time series x (t) is given as the training samples, wherein t is the time point {0, 1, 2 ...N – 1, N}, thus, a prediction for x (t) after the time point N could be gained by training the samples X (t) = {x (1), x (2)... x (N-1), x (N)}.

According to William (2007), a non-linear estimation function f(x) is listed as below:

$$f(x) = (w \cdot \phi(x)) + b,$$

wherein, "." is dot product, weight w is a normal vector, threshold b is an offset parameter, φ(x) is from a kernel function:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle.$$

A regularized risk function R_{reg} (f) is introduced to determine the parameters w and b:

$$R_{reg}(f) = R_{emp}(f) + \frac{\lambda}{2}\|w\|^2$$

wherein, λ is a capacity control factor, R_{emp} (f) is an empirical risk function:

$$R_{emp}(f) = \frac{1}{N}\sum_{i=0}^{N-1} L(x(i), y(i), f(x(i), w))$$

wherein, x (i) is a time series, i = {0, 1, 2 ...N − 1}, y (i) is the predicting results. Thus, minimizing the value of the regularized risk function R_{reg} (f) could provide optimal parameters w and b:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} L(y(i), f(x(i), w))$$

wherein, C is positive constant, L is a ϵ -insensitive loss function:

$$L(y(i), f(x(i), w)) = \begin{cases} 0 & \text{if } |y(i) - f(x(i), w)| \le \epsilon \\ |y(i) - f(x(i), w)| - \epsilon & \text{otherwise.} \end{cases}$$

wherein, ϵ is a precision constant.

Resulting from above, f(x) is calculated by the following equation:

$$f(x) = \sum_{i=1}^{N} (a_i - a_i^*) \langle x, x(i) \rangle + b$$

wherein, α_i are Lagrange multipliers.

Resulting from above, by introducing a kernel function into SVM, to map the nonlinear input space to the high-dimensional feature space Φ(x (i)), non-linear SVR is achieved. Some prevalent kernel functions are listed as below (Guangyi & Gregory, 2005):

Polynomial homogeneous function:

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j})^d$$

Polynomial inhomogeneous function:

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$$

Gaussian radial basis function:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2), \text{ for } \gamma > 0$$

Hyperbolic tangent function:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c), \text{ for some (not every) } \kappa > 0 \text{ and } c < 0$$

Radial base function:

$$k(x, \acute{x}) = exp(-\gamma \|x - \acute{x})\|^2), \text{ for } \gamma > 0$$

3.3 SVR Applications

Recently, SVM has been widely used in a variety of fields for prediction based on time series data, replacing NNs or ANNs, this is because the over-fitting problems which are the main drawbacks with NNs and ANNs in time series forecasting, can be readily solved by SVR, through the kernel functions used with, such as RBF kernel (Guangyi, & Gregory, 2005).

In relevant literatures a number of research works have been conducted to compare the performances of SVM and ANN, to determine which one is more suitable for time finance market analysis, in particular for stock market, Ajith (2007) with his colleagues conducted a large scale comparison among SVM, ANNs, DBNN (Differential Boosting Neural Network), Decision Trees, and MLP (Multiple Layer Perception), and concluded that SVM performed best, as reflected by the lower values of some prominent error metrics, including RMSE (Mean Squared Error) and MAPE (Mean Absolute Percentage Error).

The practices have proved the suitability of SVR in the field of finance market analysis, Yoo, Kim, & Jan, (2005) proposed an approach based on SVR algorithm to make trend prediction for a 7 day time period, with the data of the electricity market prices in Australian.

The results show, as compared to the use of NNs, the use of SVR can return a higher accuracy rate.

Other researchers (Kuang, et al., 2008) presented their research findings with respect to corporate credit rating forecasting in auto insurance market in Australia, they used SVM to predict the ratings, and compared the result of SVM with the results of ANNs. Again, they concluded that SVM has better performance in prediction than ANNs.

Chapter 4: Methodology

This chapter contains the information about how the research is conducted. The entire research is a quantitative research by using a quantitative comparison between the existing method and the proposed method. The following chapter firstly gives an overview of the entire method, and then depicts each component of the method. The method can be mainly divided into four parts: Data preparation, Pattern Extraction, Data Selection, and Result Prediction. The experiments setup and testing data are also described in detail in the chapter.

4.1 Method Overview

As described before, stock market is a chaotic dynamic system, stock prices vary over time as they are significantly influenced by a number of macroeconomic and microeconomic factors all the time, directly or indirectly. From the historical data of a stock observed simply extracting some rules that the stock follows, and thereby forecasting the stock movements within a certain future time period, is rather difficult and exhibits lower accuracy, as too many factors must be considered. However, while studying a stock with its graphical representation (zigzag curve), the researcher found that the similarities of the stock movements often exist (Bin, Hoi, & Gopalkrishnan, 2011). Thus, this gives a motivation for developing a geometric method to extract the movement similarities from the past data, and these similarities could be fed into the automated learning machines for training.

This section briefly introduces the method that we used in the research as shown in Figure. 3.

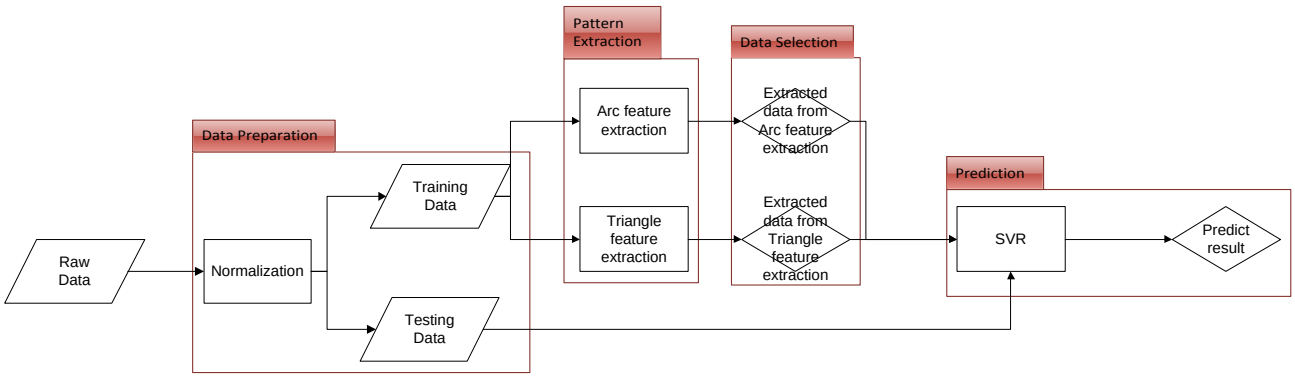


Figure 3 the flow chart of the proposed method with pattern extraction

As seen in the above figure, after the raw data is inputted into the system, it will be normalized by the data preparation component. The raw data will be divided into two parts, afterwards, the pattern extraction part including two components, Arc Pattern Extraction

and Triangle Pattern Extraction, will be conducted to selected appropriate training data. Next, the selected training data will be utilized to construct a training model, and then the first testing data will be input into the model for regression. Lastly, after the testing data has been labeled, it will be used as a new training data for the next prediction process. The entire process can be regarded an iterative process for predicting the rest of testing data.

4.2 Differences from the existing method

As known, the topic about regression/prediction in stock market has been popular for many years. During the last decades, with the development of data mining and machine learning, more and more advanced techniques has been introduced into this field to enhanced the performance of prediction. Using SVR as a regression tool is not a very new topic, since the strong ability of SVM has been recognized by most researchers and practitioners. The following figure shows a typical method for regression by employing SVR.

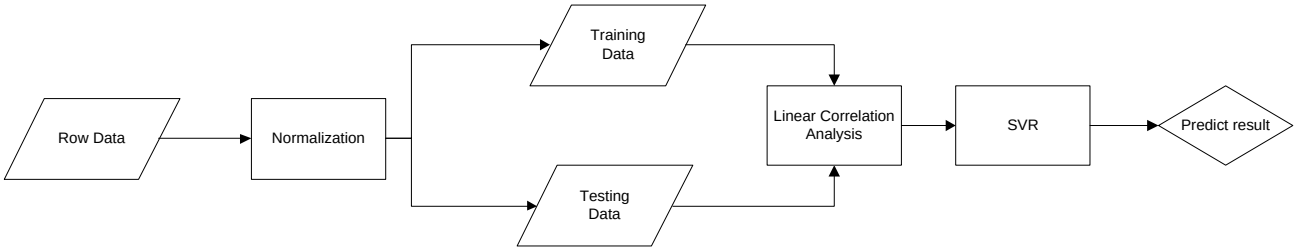


Figure 4 a typical regression method without pattern extraction

Another advanced method for prediction is use the correlation extraction to enhance accuracy.

The main differences between the existing method without pattern extraction outlined on Figure 4 above and the proposed method have been listed as below.

1) It clearly appears that the proposed method uses geometric methods for pattern extraction before conducting the prediction, which can highly improve the performance of the prediction.
2) The existing method has employed a linear correlation analysis before the data being inputted into the SVR model for prediction, but it fails in performance, since the stock data is often stochastic and non-linear distributed. A linear model is not capable of handling the non-linear problems.

In the following text, we will use “without pattern extraction” to represent the existing method. In addition, “with pattern extraction” will be used to represent the proposed method.

4.3 The proposed method

The section introduces the proposed method in detail. Each four component will be depicted step by step in the following text.

The core idea of our method is 1) selecting the first testing instance for prediction; 2) combining the selected testing data with its previous 19 days training data to form a 20 days' time window (here we define the time window as "Object Time Windows"); 3) shifting the time window in a backward direction until all training data has been covered; 4) during the shifting process, we intend to find the trend trajectories formed by historical data which are similar to the "Object Time Windows"; 5) selecting the historical data that satisfies condition presented in the above step to form a training data model by SVM; 6) using the model to predict the label of the first selected testing instance. Up to now, the prediction process is completed.

4.3.1 Data Representation

For prediction in stock markets, we have selected several stock indexes as the experimental data. An index (or the closing prices of a stock) can be represented as a vector Y as follows:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \end{bmatrix}, t = 1,2,...,T$$

where y_1 denotes the closing price of the specific time point t, T represents the time frame of the stock that we selected.

4.3.2 Data Preparation

In this part of the method, we will conduct two steps, the first is normalization, and the second step is data separation.

4.3.2.1 Normalization

Normalization is a very important step in data mining area, which aims to regularize the raw data into a specific scale. In our case, we normalized the data into the scale [0, 1]. Normally, there are several approaches to complete the normalization, a common way is to use the

statistical standardize method. However, in our case, this is not an appropriate way, since such type of standardization usually normalizes the data into a scale of $[-K, K]$ which does not satisfies our demands.

According to Bulcock(1980), normalization is a key stage for regression. He introduced a more suitable for normalization stock prices for regression, the formula that we used has been listed as below:

$$Y^N = \frac{Y}{\max(Y)} = \begin{bmatrix} y_1^N \\ y_2^N \\ \vdots \\ y_t^N \end{bmatrix}$$

where Y^N denotes the normalized stock price vector, $\max(Y)$ calculated the maximum value in vector Y.

4.3.2.2 Data Separation

In fact, it is obvious to see that a classification problem can be regarded as a regression problem and vice versa. Thus, as usual, we will separate the data into two parts, training data and testing data, here in our case; we selected 9:1 as the ratio of the training data to the testing data. For example, if we select 10 years historical data, we will use the previous 9 years data as training data, and the last 1 year data as testing data, as show in Figure 5.

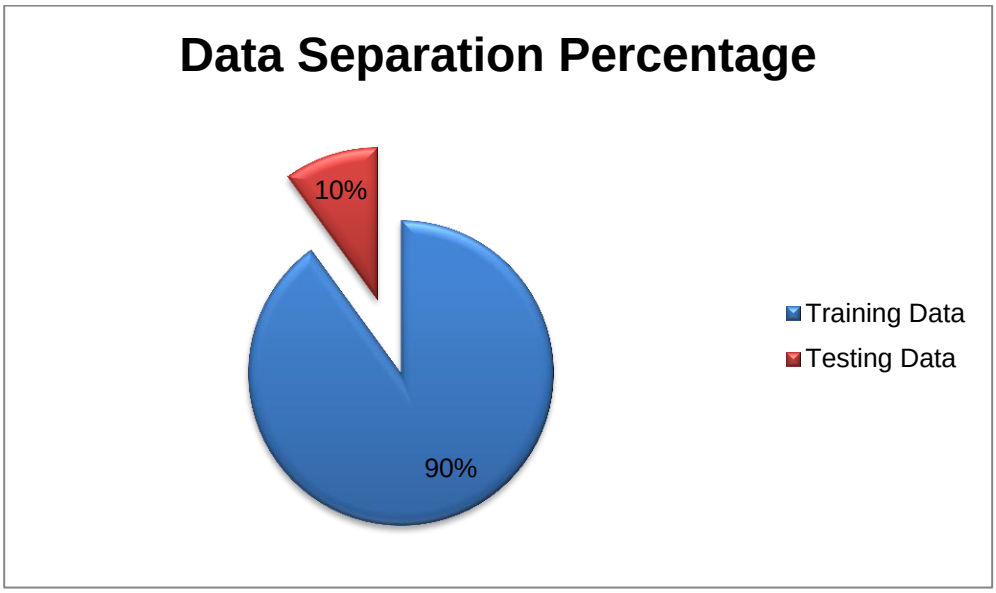


Figure 5 Data Separation Percentage

Note that, after the data separation, merely training data will be used and the testing data will never be touched during the model training process, only if the testing process starts, the first test data will be selected for testing/prediction.

4.3.3 Pattern Extraction

4.3.3.1 Data Representation

Data often has two types of representations, the numerical representation and the graphical representation. For example, a typical numerical representation is given as below:

Date	Open	High	Low	Close	Volume	Adj Close
3/01/2008	4041.38	4050.17	4031.56	4033.93	18106800	4033.93
4/01/2008	4033.48	4033.48	4004.14	4010.82	13081200	4010.82
7/01/2008	4010.82	4010.82	3950.35	3953.58	14626000	3953.58
8/01/2008	3953.58	3955.32	3928.4	3928.4	23311800	3928.4
9/01/2008	3928.4	3928.4	3883.76	3912.17	25542200	3912.17
10/01/2008	3912.17	3914.99	3891.91	3899.88	20254600	3899.88
11/01/2008	3899.88	3913.07	3869.98	3872.18	19247800	3872.18
14/01/2008	3872.18	3872.18	3812.1	3824.21	26525800	3824.21

Table 1 Numerical Data Representation

As seen from Table.1, the numerical data representation is often used for calculation and analysis. A more intuitive way for analysis mostly is to use a graphical representation for the data. For example, the following figure shows a graphical representation.

Figure 6 Data Graphical Representation

In figure 6, besides the exact price value in a time frame, it also shows a decreasing trend of the stock prices. This will greatly help us to analyze the pattern of the stock.

Our idea is derived from the graphical representation; we believe that there must be a graphical pattern existing in a stock, and a correlation existing between the historical data and the present data as well. Thus, we intend to extract patterns of the stock price data from the graphical trend for prediction. In the following text, we will depict two graphical/geometric methods employed in our research.

4.3.3.2 Correlation in graphical representations

Correlation, in statistics, is defined as the relationship between two variables. Generally, there exist many types of correlations, such as distance correlation (Székely, Rizzo, & Bakirov, 2008), Pearson's linear correlation (Rodgers & Nicewander, 1988) and Brownian correlation (Székely & Rizzo, 2009).

In our case, we defined another type of correlation existing in the graphical figures. Similarity, to some extent, can be regarded as a type of correlation. Based on the stock data, we are able to distribute all data in a 2-D plane with axis X for time instance and axis Y for price value. As seen from Figure.6, those data points can be connected together to form a trend trajectory. Our aim is to find the trajectories similar to the targeted trajectory formed by one target sample for prediction with its previous 19-day data samples in a 20-day time window. However, those trajectories normally are curly in the plane, which are regarded as non-linear. As known, it is difficult to conduct analysis directly on a non-linear problem, especially on a non-linear problem being so irregular. Therefore, we applied two pattern extraction methods to solve the issue as depicted in the following sections. In addition, after pattern being extracted, we adopted two approaches to measure the similarity (i.e., the correlation), which will be detailed in section 4.3.3.5.

4.3.3.3 Triangle pattern extraction method

The first graphical pattern extraction method we used in this research is triangle pattern extraction method. As we have discussed previously, it is a difficult task to extract pattern from the irregular ZigZag trends directly portrayed by the stock price data. Thus, we proposed the triangle pattern extraction method to simplify the trend by two linear functions.

The method is to construct a linear model with application of two linear functions, which can be used to solve a non-linear problem for stock market prediction, instead of a simple linear regression model. The linear model can be constructed according to the following steps:

1) Select a subsequent 20 samples to form 20-day time window shown below (the hat

"^" means that it is only partial data rather than the entire dataset):

$$\hat{Y}^N = [y_1^N, y_2^N, ..., y_{20}^N]^T$$

2) At this stage, we could obtain a graphical representation of the selected data:

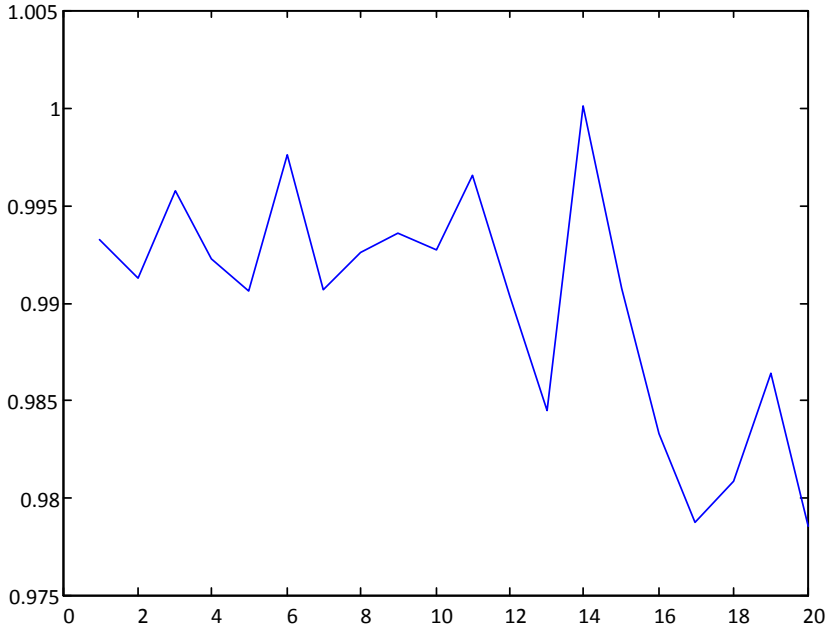


Figure 7 the stock price graphical representation in a 20-day time window

3) Separating the data into two vectors with size 10×1 evenly as follows:

$$\hat{Y}_1^N = [y_1^N, y_2^N, ..., y_{10}^N]^T$$
$$\hat{Y}_2^N = [y_{11}^N, y_{12}^N, ..., y_{20}^N]^T$$

4) Choosing the minimum data and the maximum data from the two parts respectively to obtain 4 values, which can be formed to 4 coordinates in a 2D plane as:

Point 1: $(t_1, \max(\hat{Y}_1^N))$, Point 2: $(t_2, \min(\hat{Y}_2^N))$

Point 3: $(t_3, \min(\hat{Y}_1^N))$, Point 4: $(t_4, \max(\hat{Y}_2^N))$

5) At this stage, graphically, we could separate the graph into two parts, and indicate the 4 points on the graph as below:

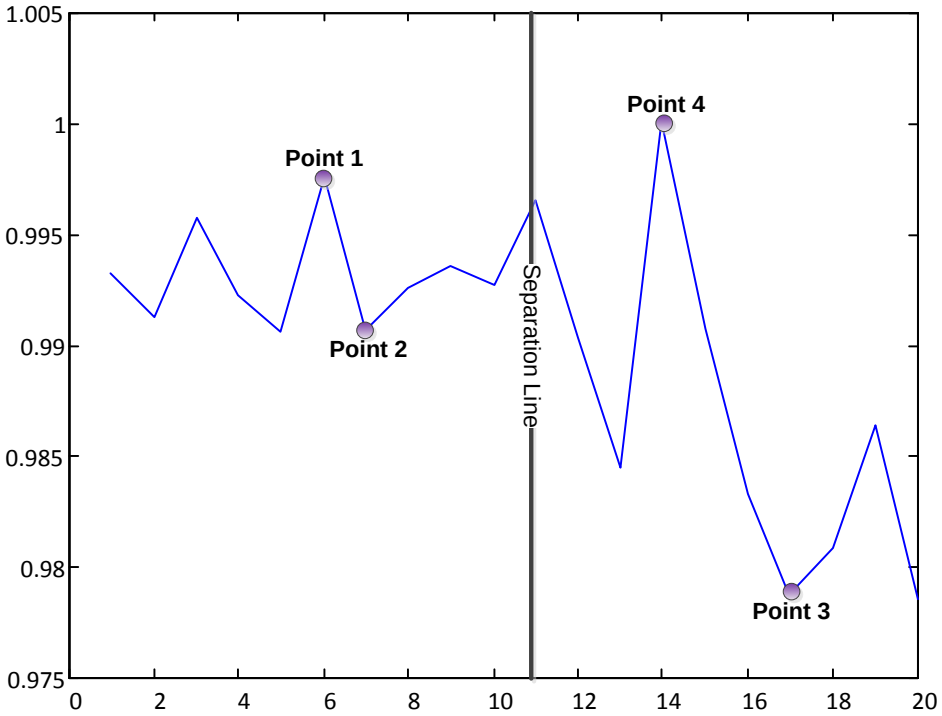


Figure 8 the stock price graphical representation in a 20-day time window with 4 highlighted points

6) Using the point 1 and point 2 to obtain the first linear line, using point 3 and point 4 to obtain the second line. Hence, we can use two linear functions to obtain the model as:

$$\lambda_1 = a_1\omega_1 + b_1$$
$$\lambda_2 = a_2\omega_2 + b_2$$

Substituting the points into the above equation, we can obtain the linear model for the research as:

$$\lambda_1 = \frac{\max(\hat{Y}_1^N) - \min(\hat{Y}_2^N)}{t_1 - t_2}\omega_1 + \frac{t_1 \times \min(\hat{Y}_2^N) - t_2 \times \max(\hat{Y}_1^N)}{t_1 - t_2}$$
$$\lambda_1 = \frac{\min(\hat{Y}_1^N) - \max(\hat{Y}_2^N)}{t_3 - t_4}\omega_1 + \frac{t_3 \times \max(\hat{Y}_2^N) - t_4 \times \min(\hat{Y}_1^N)}{t_3 - t_4}$$

7) At this stage, we can also connect point1 with point3, and connect point2 with point4.Thus, two triangles can be obtained as shown in the figure 9:

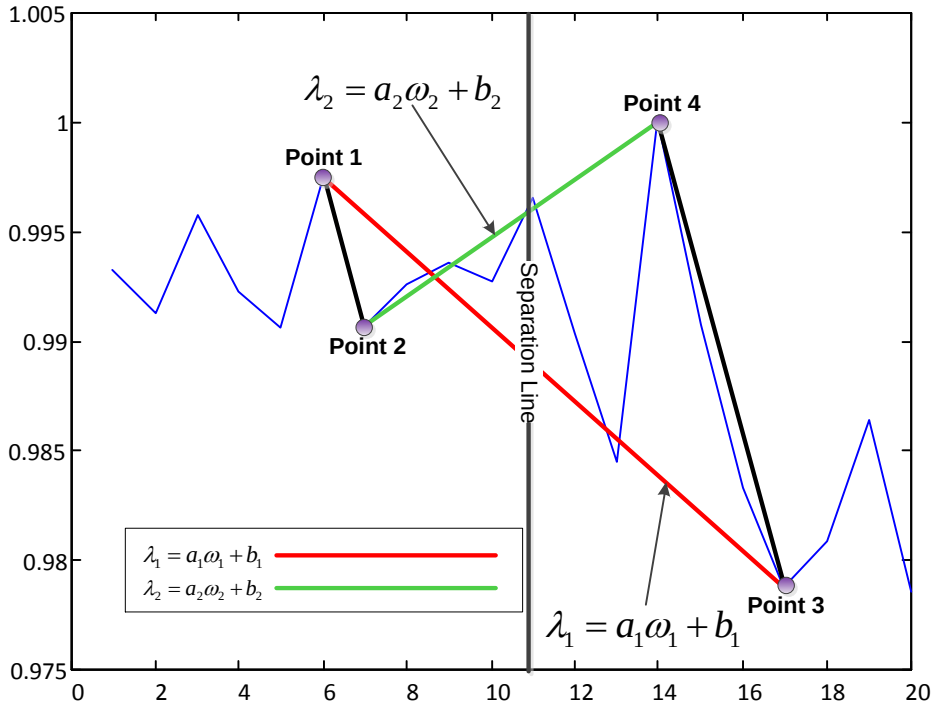


Figure 9 Triangle Patterns in the given 20-day stock price data

Consequently, we extracted the triangle pattern as defined by points 1, 3 and 4 of the stock price data in this 20-day time window. In this way, we could extract many triangle patterns from the shifting time window iteratively.

By using this method, we are able to compare the patterns among different time windows. For example, if we assume point 3 is the selecting data, we can extract the pattern in this time-window as shown in Figure.10 (a). Next step is to use the time window shifted backwards in the historical data to find the similar pattern as described in Figure.10 (b).

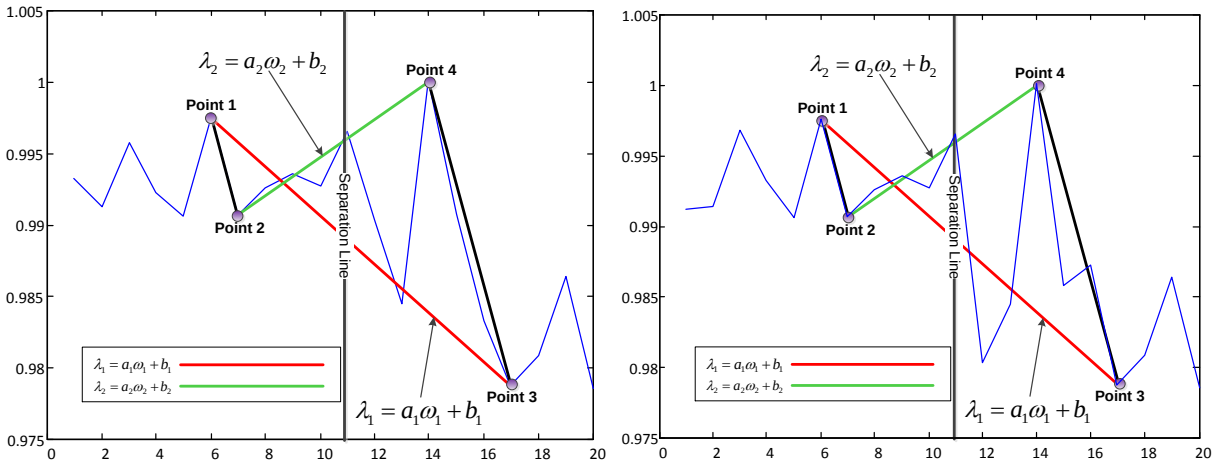


Figure 10 extracting similar triangle patterns in the historical data

As seen from the figure, we can see that these two patterns are very similar. However, this is just from a graphical way to determine the similarity. It is necessary to demonstrate the similarity between these two patterns. To accomplish the task, we could use correlation (similarity) measurement defined in the following section 4.3.3.5. After this type pattern extraction (Triangle Pattern Extraction) and correlation measuring, those data will be marked for data selection.

4.3.3.4 Arc pattern extraction method

In fact, based on the above triangle method, we have already been able to extract a series of similar patterns of the stock prices to the target time window containing the first testing data. However, during the experiment, we found that if only use the method, the selected training data is very insufficient. This may be because the triangle method we used is a linear method which may not be suitable to handle a non-linear problem like this. Thus, to overcome the deficiency of the linear method, we have adopted a method firstly proposed by Song (2010), the Arc pattern extraction method. We believe that based on both methods, the sufficiency of the training data will be satisfied and the performance of prediction will be improved.

According to Song (2010), the pattern of the stock trend can be described as Arcs. As discussed previously, the trend can also be graphically described as Triangles, but it ignores some small variations of the trend. In contrast to the triangle method, the Arc pattern extraction can handle those small variation or fluctuations very well.

The method adopts an arc, which can represent the trend and is obtained by cutting from the circle according to the appropriate angles, to form a non-linear model. The non-linear model has been presented as below (Song, 2010):

$$(x - x_0)^2 + (y - y_0)^2 = R^2 \quad \left| \begin{array}{l} x_0 = 0, y_0 = R \\ x \in [0, sin\alpha \cdot R\sqrt{2(1 - cos2\alpha)}] \end{array} \right.$$

where $\alpha \in [0, \frac{\pi}{4}]$ is a parameter describing the increasing or decreasing speed.

In this method, the most difficult step is to determine the center of the circle, the angle α , and the radius R . In our case, this step can be done automatically by the method of exhaustion. The detailed step will be listed as below:

1) The coordinators of the starting point and the end point can be described as:

Start point: (t_1, y_1^N), End point: (t_{20}, y_{20}^N)

2) Highlighted the starting point and the end point of the selected data as shown in the figure:

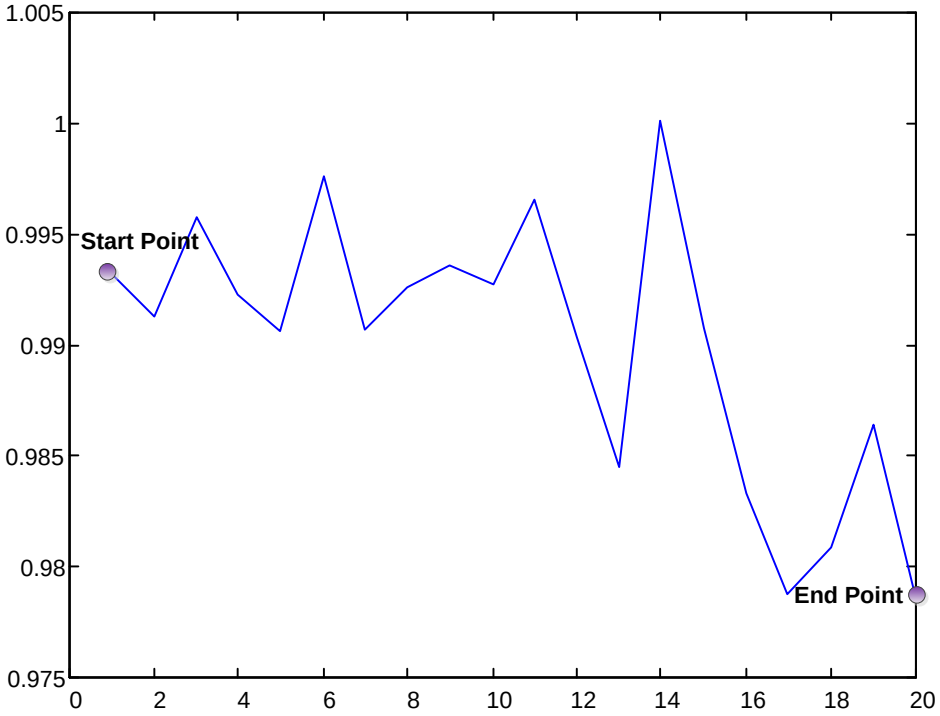


Figure 11 a stock price trend with the highlighted start point and end point

3) Connect a line between the two points, the line function can be described as:

$$\lambda = a\omega + b$$

4) Substituting the coordinates of the two points (start and end) into the above function, we can obtain as :

$$\lambda = \frac{y_1^N - y_{20}^N}{t_1 - t_{20}}\omega + \frac{t_1 \times y_{20}^N - t_2 \times y_1^N}{t_1 - t_{20}}$$

5) According to the above function, we could determine the midnormal on the line, the midnormal function can be described as follow:

$$\lambda_m = a_m\omega_m + b_m$$

6) The two lines can be described graphically as the following figure:

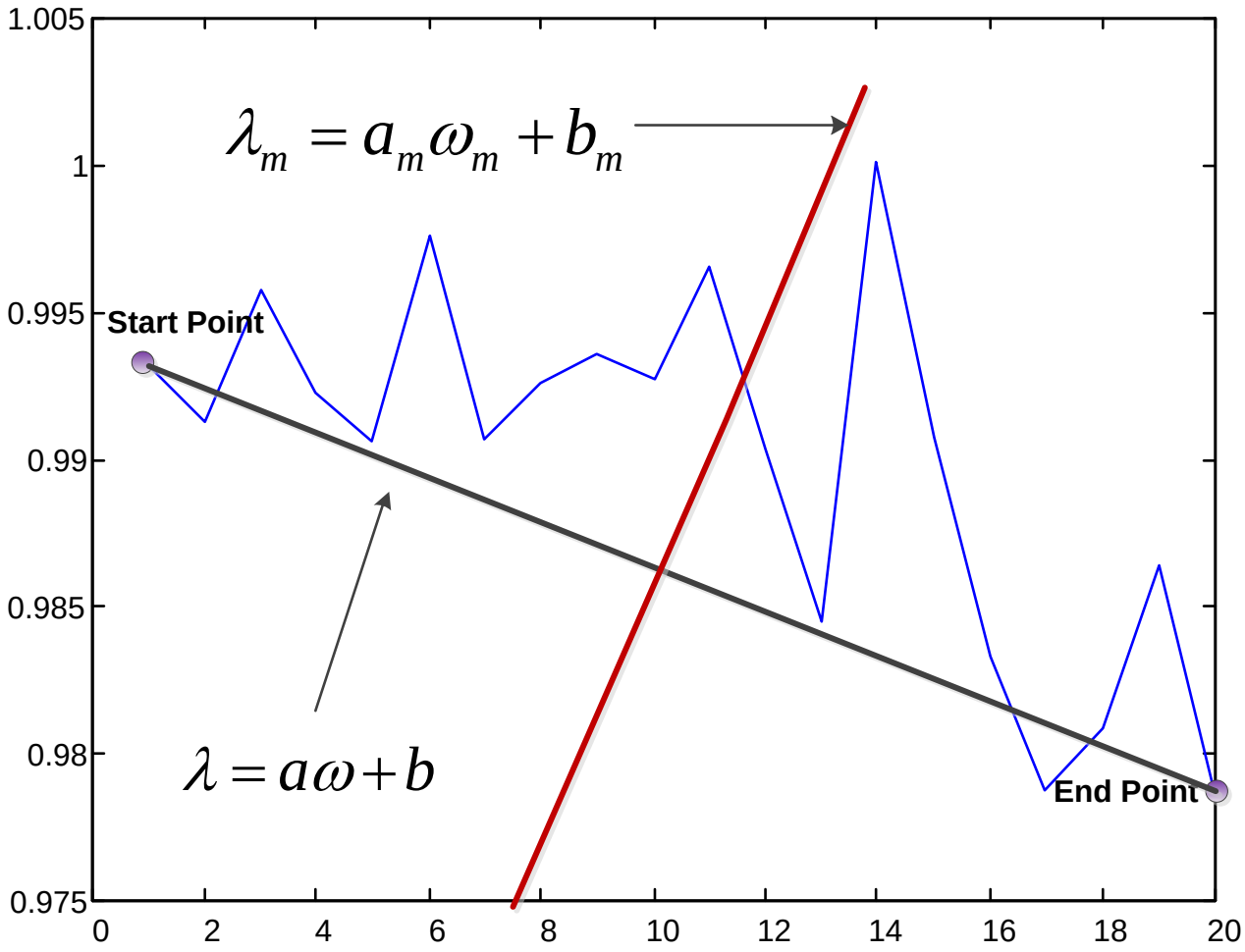


Figure 12 the connecting line between the start point and the end point, with its midnormal

7) In this stage, we will use each point on the midmal as a circle center, and the start point as well as the end point as two points on the circle. Iteratively, we could obtain a group of circles. Then, we will use the following equation to verify the distances between the arc of the circle and the points on the ZigZag trend. The equation is given as below:

$$d_{euc}(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2},$$

8) Lastly, we will choose the minimum distance, and select that arc as the arc to represent the trend as shown in the figure.

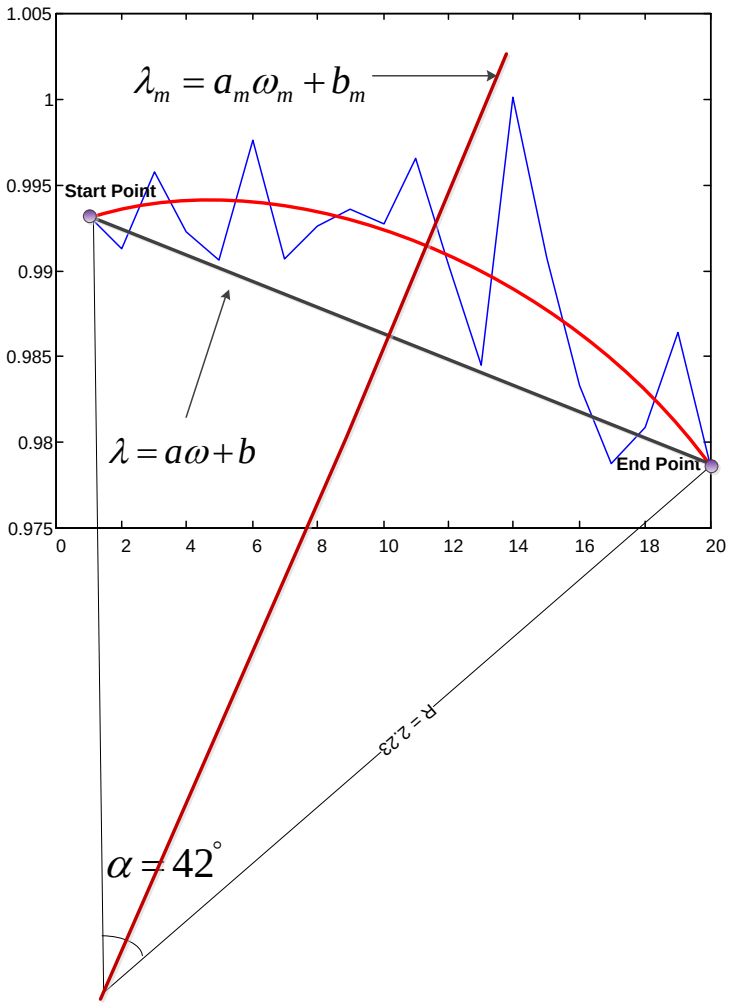
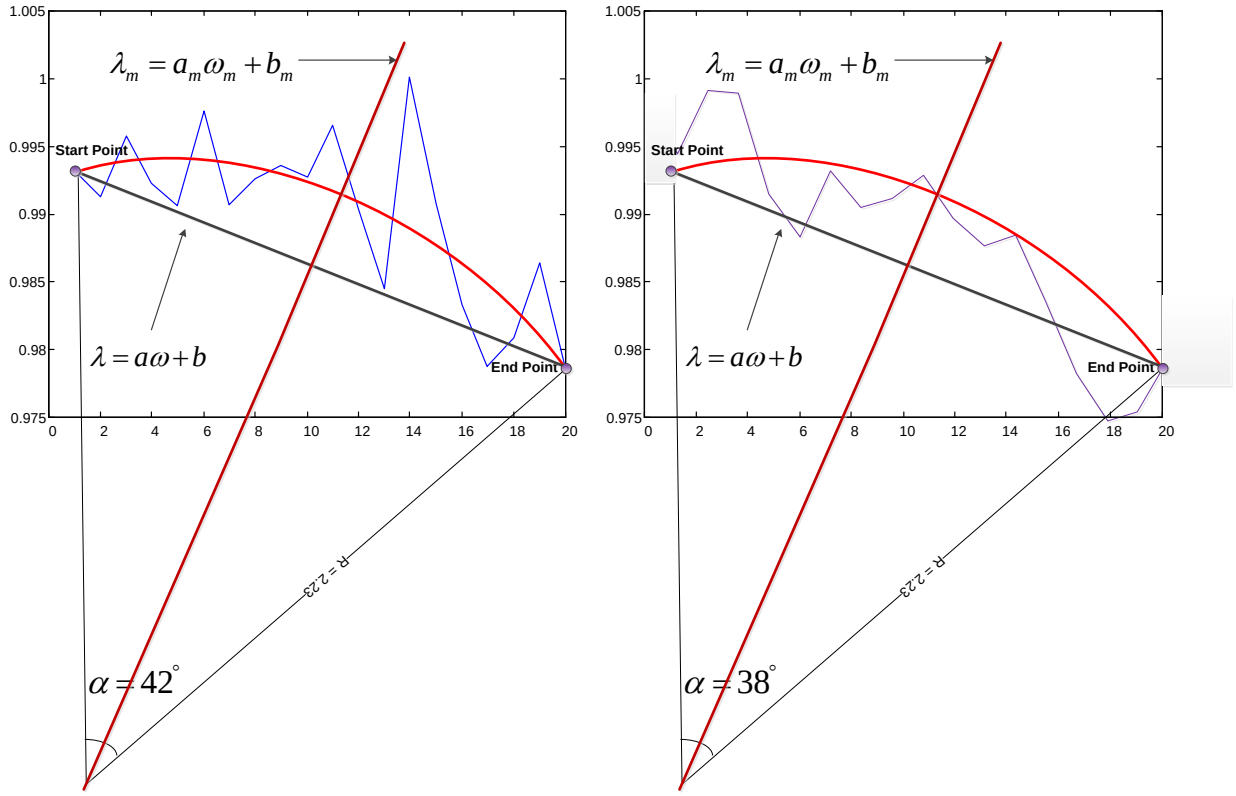


Figure 13 the arc pattern extraction

Based on the above equation, we can extract the Arc pattern of the data in this 20-day time windows.

Lastly, we can adopt the same similar pattern searching approach mentioned in the above section 4.3.3.2 to find the similar patterns in the historical data. The following figures show a similar arc pattern between two different time windows.



(a) (b)

Figure 14 extracting similar triangle patterns in the historical data

Figure 14 shows that the two arc patterns are very similar. Similar to the above triangle method, we must conduct a mathematical way to calculate the similarity according to the approach depicted in section 4.3.3.5.

4.3.3.5 Methods Combination

As we mentioned above, the two pattern extraction methods are both used for extracting patterns. The reason of using two methods together is that, after the linear method – triangle method – is used, sometimes only a few training data samples can be selected for next step to train the model. To solve the problem, we introduced the second method – Arc method. In this way, the sufficiency of the training data can be guaranteed.

As we discussed above, if only one method has been adopted, it may causes training data insufficiencies. It is a challenge to define how much training data is sufficient. Using an automatic way is not an easy task to accomplish, thus we use a manual way to achieve the final goal, we setup a threshold to check whether the training data is sufficient. We defined that 90% data should be used as traing data, and 10% is used for testing data, so we select 9 to 1 as the rate between training data and testing data. This rate is obtained by

many times repeated experiments, since we found that, in so many experiments, the accuracy is the highest when this rate is employed.

According to the above statements, the methods are actually applied in a parallel and independent way; patterns extracted from the methods are both adopted. However, there must be some overlaps, in other words, redundant data. Another process will be adopted to filter out those redundant data before the training data is actually used for modeling; this has been indicated in the section 4.3.4 as well. The process is very simple, we can use time as reference to filter out those data who is at the same time point with the same price value.

4.3.3.6 Correlation Measurement

As discussed in the above sections, we have already conducted two pattern extraction methods. Usually, to measure similarity, we could adopt the distance between two variables. The easiest way is to conduct Euclidean distance between variables as shown in the below equation:

$$d = \frac{\sum_{i=1}^{20} \left\| \hat{y}^{N}{}_{i} - y^{N}{}_{i} \right\|}{20}$$

Where d denotes the distance between two variables, 20 represents the time window, and $\hat{y}^{N}{}_{i}$ represents the target time window after a parallel move, $y^{N}{}_{i}$ means the previous time window.

Though we have the same time window of 20 days, and the window will shift backward from the end, for the two comparing time windows, we have made an assumption that they are at the same point, then these two patterns are in the same calculation level. At then, the similarity can be measured.

Afterwards, we setup a threshold; only when $d < 0.01$ will be regarded as similar trends. Actually, the threshold is defined manually through the software, so users can input it according to different situation. In our case, we suppose that when $d = 0$ the two patterns are the same, so we select those distances which are very close to zero.

At last, if the threshold cannot be satisfied by all data, we will increase the threshold. In fact, for our experimental data, the threshold works well. At this stage, we can extract several data forming those similar patterns for data selection step.

4.3.4 Data Selection

As we have already extracted those similar patterns from the historical data, the next step requires to be done is to select the corresponding data. Since we started from the graphical representation of the data, the current stage is to utilize the numerical representation of those selected data. The data has only one feature which is the normalized price value, as described below:

$$Y_1^{\,N} = \begin{bmatrix} y_{t_1} \\ \vdots \\ y_{t_2} \end{bmatrix}, \quad Y_2^{\,N} = \begin{bmatrix} y_{\tilde{t}_1} \\ \vdots \\ y_{\tilde{t}_2} \end{bmatrix} \quad \ldots \quad Y_p^{\,N} = \begin{bmatrix} y_{\hat{t}_1} \\ \vdots \\ y_{\hat{t}_2} \end{bmatrix}$$

Because the pattern similarity cannot be guarantee to be a continuous way, so the data would be selected as above in several different blocks according to different time frame. In addition, sine we were using the time window shifting from the end to the start point, thus there may exist redundant data.

Consequently, the last step of data selection is to remove redundant data and combine all training data vectors into one vector. The data vector will be inputted into the SVR for next step prediction.

4.3.5 Prediction

According to the training data selected from the above step, we can input those data into the SVM to construct a training model. After the model has been trained, we could put the target data instance into the training model for prediction. As this stage is just to use an existing method to conduct research, the details of training and testing has not been detailed in the thesis.

4.3.6 Summary

The proposed method is a hybrid based on the above two graphical pattern extraction approaches. We use the two pattern extraction method as a complimentary method, since the triangle method is a type of linear method which may not be able to handle the non-linear problem very well. At this time, the Arc patter as a simple non-linear approach can be applied to handle the problem and further improve the effectiveness of data selection. Thus, the performance of prediction can be improved by applying the two complementary

methods. Some characteristics of the two methods have been discussed in the following paragraphs:

1) The triangle method is appropriate for handling the problem where the prices trend varies significantly in a short time. In addition, it calculates only two linear functions, which is much more efficient.
2) For the Arc method, it is able to handle most cases; especially it uses a non-linear way to selected data, which is closer to the real trend than the linear lines.
3) Two methods are complementary, sometimes it may not be possible to extract a similar Triangle pattern in the historical data. For this case, the Arc method can be used in a complementary way.
4) Till now, we haven't met such a situation that both methods are not able to extract patterns from the historical data. However, to be a robust system, such a situation should be considered. In the future, we could drill into the problem to improve the robustness of the system.

The key improvement of the correlation extraction method proposed is the graphical trend similarity approximation, using arcs and triangles to represent the trend of a portion of the whole time series; it is called geometric correlation extraction method in this research. Within the conventional correlation extraction methods, the evaluation of the trend similarity between two time series is carried out by calculate the distance of the two time series to the time series observed. However, with the pattern extraction methods proposed in this research, the evaluation of the trend similarities is turned to calculate the distance of the arcs or triangles. This means, the new method extracts correlation from general variation trends, avoids extracting correlation form the time series directly, the mismatch problem is thus solved.

In addition to above, weighted Pearson's correlation algorithm could be used to measure the strength of the correlations (Hong, 2009), thus the correlations with other factors, such as with correlative stocks or macroeconomic data, are introduced to assist the trend prediction, and have different weights. Therefore, the correlation method proposed is able to realize the research goal, which is to combine the technical analysis and fundamental analysis, concerning both microeconomic and macroeconomic factors, for enhanced market trend forecast.

The prediction implemented in this research is based on the proposed geometric method and SVR. The data required in this research project is stock prices over time and collected

from electronic resources. The proposed method will be examined by comparing its results with the results of SVR on its own, thus, several accuracy metrics are used to measure the results.

4.4 Experiment Setup and Data Used

This research project includes an experiment for verifying the method proposed, the experiment is implemented in the MATLAB environment. In the experiment, the proposed prediction method is performed based on SVR RBF kernel; the parameters for SVR machine learning are set by MATLAB default.

The data required by the experiment is stock market data, which is time series data in nature. In detail, the data include the stock indexes of five counties or regions from 2004 to 2010: NZX 50 index in New Zealand, HangSeng index in Hongkong, S&P 500 index in U.S., Nikkei index in Japan, and All Ordinaries index in Australia, collected from Yahoo Finance. The following table shows an example of the data format, wherein only the closing price is used in prediction.

Date	Open	High	Low	Close	Volume	Adj Close
2010-12-31	3334.27	3334.50	3308.68	3309.03	8455000	3309.03
2010-12-30	3325.62	3334.34	3319.79	3334.27	9556000	3334.27
2010-12-29	3329.21	3329.21	3317.40	3325.62	8355800	3325.62
2010-12-24	3333.76	3338.87	3325.04	3329.21	22695000	3329.21

Table 2 Dataset Description

Furthermore, in the experiment, the predictions respectively for the time horizons 1-5 days for each stock index are performed with the proposed method and SVR, and with SVR only, and the results are measured by some accuracy metrics for comparison.

The experiment is conducted on an Intel i5 core 1.87 computer with the software MATLAB 7.0 based on data from ALL ORDINARIES; HangSeng; Nikkei 225; NZ50; S&P 500

INDEXs from 2nd January 2008 to 31st December, 2010, we created a comparison of regression between regression with pattern Extraction and without With Pattern Extraction. Figure.6 shows the interface of programming on proposed method.

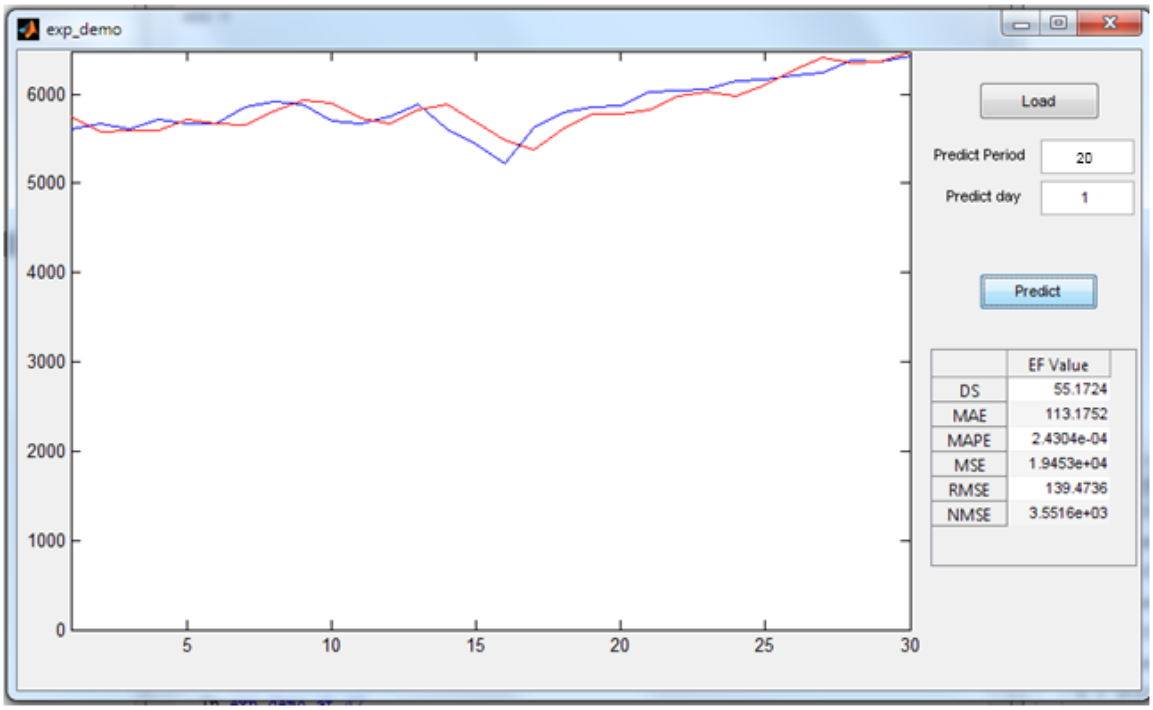


Figure 15 the screen shot of the prototype with the proposed method (coding in Matlab)

4.5 Performance Metrics for Time Series Prediction

To compare the performance of the different algorithms in stock market forecasting based on the proposed geographic method in terms of accuracy, some error metrics are involved. Generally, the measurements of the machine learning/ Data mining are often accuracy, speed, and memory usage. However, the measurement of the forecasting prediction often choose standard error measures (Kolarik & Rudorfer, 1994), including Directional Symmetry (DS), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Squared Error (RMSE), and Normalized Mean Square Error (NMSE). Those measures are suitable for comparing time series with different characteristics, all of them are widely used due to their own special characteristics, and useful for comparing the same data but tested on different methods.

4.5.1 DS

Directional Symmetry (DS) is a common means used to measure the performance of the machine prediction; it uses the following equation to calculate the DS coefficient:

$$\mathrm{DS}(t,\hat{t}) = \frac{100}{n-1}\sum_{i=2}^{n-1} d_i,$$
$$d_i = \begin{cases} 1, & \text{if } (t_i - t_{i-1})(\hat{t}_i - \hat{t}_{i-1}) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

wherein, d is the direction, t is the actual value and $\hat{t}$ is the predicted value. Higher values of the DS metric indicate prediction trends follow the actual trends and thus higher values indicate better performance.

4.5.2 MAE

The Mean Absolute Error (MAE) is the average of the absolute error metric, simply measuring the distance between the actual signal and the predicted signal, it is calculated by:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|$$

where is the predicted value and y_i is the true value, and $|f_i - y_i|$ is the absolute error.

4.5.3 MSE

The Mean Square Error (MSE) is an estimation function T for an unknown parameter θ, thus, MSE can be calculated by:

$$\mathrm{MSE}(T) = \mathrm{E}((T-\theta)^2)$$

or

$$\mathrm{MSE}(T) = \mathrm{var}(T) + (\mathrm{bias}(T))^2$$

wherein, var is the variance, and bias is the bias between T and θ.

4.5.4 RMSE

The Root Mean Squared Error (RMSE) is the root of MSE:

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})}$$

wherein, $\hat{\theta}$ is the estimation function with respect to the unknown parameter θ. It is another common metric for estimating the deviations in prediction. The RMSE gives more weight to larger errors unlike MAE that weights all errors the same.

4.5.5 NMSE

Normalized Mean Square Error (NMSE), or Standard Error, is a normalization of MSE, for estimating the overall deviations. It is defined as:

$$NMSE = \frac{1}{N}\sum_{i}\frac{(P_i - M_i)^2}{\overline{P}\,\overline{M}}$$

wherein,

$$\overline{P} = \frac{1}{N}\sum_{i} P_i$$

$$\overline{M} = \frac{1}{N}\sum_{i} M_i$$

These error metrics are used in the experiment to examine the performances of two predictions (SVR with geometric and SVR on its own).

Chapter 5: Experimental Results

5.1 Introduction

In the experiment, the historical data of five stock indexes have been respectively learned by SVR machine to evaluate the extracted correlation knowledge by the proposed geometric method. This chapter outlines and compares the results from the experiment.

5.2 Results

According to literature, we have selected directional symmetry (DS), mean squared error (MSE), root mean squared error (RMSE), normalized mean square error (NMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) to evaluate the performance of proposed method on five scenarios. Each scenario is designed to predict the value of the stock closing price for the following days, as day 1, day 2, day 3, day 4 and day 5. Table 2 depicts the results obtained.

Prediction of day 1	Method	DS	MAE	MAPE	MSE	RMSE	NMSE
ALL_ORDINARIES_closingPrice	With Pattern Extraction	55.1724	113.1752	2.4304e-04	1.9453+04	139.4736	3.5516e+03
	Without Pattern Extraction	55.1724	112.5912	2.43E-04	1.92E+04	138.6536	3.51E+03
HangSeng_closingPrice	With Pattern Extraction	34.4828	945.7056	5.01E-04	1.44E+06	1.20E+03	2.63E+05
	Without Pattern Extraction	34.4828	950.9191	5.01E-04	1.45E+06	1.20E+03	2.64E+05
Nikkei225_closingPrice	With Pattern Extraction	51.7241	315.39	2.21E-04	1.30E+05	360.3047	2.37E+04
	Without Pattern Extraction	48.2759	3.17E+02	2.21E-04	1.31E+05	3.62E+02	2.40E+04
NZ50_clos	With Pattern	65.5172	36.6724	4.00E-04	1.88E+03	43.3994	343.8794

ingPrice	Extraction						
	Without Pattern Extraction	65.5172	4.05E+01	4.00E-04	2.22E+03	4.71E+01	4.06E+02
S&P500_INDEX_closingPrice	With Pattern Extraction	55.1724	18.0269	1.12E-04	515.53	22.7053	94.1225
	Without Pattern Extraction	55.1724	1.80E+01	1.12E-04	5.13E+02	2.26E+01	9.36E+01

Prediction of day 2	Method	DS	MAE	MAPE	MSE	RMSE	NMSE
ALL_ORDINARIES_closingPrice	With Pattern Extraction	58.6207	163.9008	1.11E-04	4.07E+04	201.7916	7.43E+03
	Without Pattern Extraction	58.6207	164.9078	1.11E-04	4.09E+04	202.2338	7.47E+03
HangSeng_closingPrice	With Pattern Extraction	41.3793	1.09E+03	6.60E-04	1.97E+06	1.40E+03	3.60E+05
	Without Pattern Extraction	41.3793	1.10E+03	6.60E-04	1.99E+06	1.41E+03	3.62E+05
Nikkei225_closingPrice	With Pattern Extraction	55.1724	377.9667	3.78E-04	2.43E+05	493.0015	4.44E+04
	Without Pattern Extraction	51.7241	3.83E+02	3.78E-04	2.45E+05	4.95E+02	4.47E+04
NZ50_closingPrice	With Pattern Extraction	65.5172	62.6138	7.65E-04	5.46E+03	73.8697	996.2594

	Without Pattern Extraction	68.9655	6.44E+01	7.65E-04	5.69E+03	7.54E+01	1.04E+03
S&P500_INDEX_closingPrice	With Pattern Extraction	48.2759	24.2318	5.90E-04	1.06E+03	32.5415	193.3365
	Without Pattern Extraction	48.2759	2.42E+01	5.90E-04	1.06E+03	3.25E+01	1.93E+02

Prediction of day 3	Method	DS	MAE	MAPE	MSE	RMSE	NMSE
ALL_ORDINARIES_closingPrice	With Pattern Extraction	51.7241	197.8694	5.81E-04	5.70E+04	238.7153	1.04E+04
	Without Pattern Extraction	51.7241	198.1647	5.81E-04	5.74E+04	239.5952	1.05E+04
HangSeng_closingPrice	With Pattern Extraction	41.3793	1.49E+03	0.0011	3.41E+06	1.85E+03	6.22E+05
	Without Pattern Extraction	41.3793	1.47E+03	1.10E-03	3.37E+06	1.83E+03	6.15E+05
Nikkei225_closingPrice	With Pattern Extraction	44.8276	511.261	9.66E-04	3.75E+05	612.4582	6.85E+04
	Without Pattern Extraction	37.931	5.01E+02	9.66E-04	3.70E+05	6.08E+02	6.76E+04
NZ50_closingPrice	With Pattern Extraction	48.2759	80.333	0.001	9.80E+03	98.9811	1.79E+03

	Without Pattern Extraction	51.7241	8.44E+01	1.00E-03	1.03E+04	1.01E+02	1.88E+03
S&P500_INDEX_closingPrice	With Pattern Extraction	44.8276	30.7888	7.08E-04	1.53E+03	39.1589	279.9626
	Without Pattern Extraction	44.8276	3.04E+01	7.08E-04	1.51E+03	3.89E+01	2.76E+02

Prediction of day 4	Method	DS	MAE	MAPE	MSE	RMSE	NMSE
ALL_ORDINARIES_closingPrice	With Pattern Extraction	72.4138	209.4398	3.23E-04	6.78E+04	260.4761	1.24E+04
	Without Pattern Extraction	51.7241	197.8694	5.81E-04	5.70E+04	238.7153	1.04E+04
HangSeng_closingPrice	With Pattern Extraction	58.6207	1.53E+03	0.0032	3.26E+06	1.80E+03	5.95E+05
	Without Pattern Extraction	58.6207	1.53E+03	3.20E-03	3.26E+06	1.80E+03	5.95E+05
Nikkei225_closingPrice	With Pattern Extraction	55.1724	588.2859	2.55E-04	4.84E+05	695.8743	8.84E+04
	Without Pattern Extraction	51.7241	5.89E+02	2.55E-04	4.87E+05	6.98E+02	8.89E+04
NZ50_closingPrice	With Pattern Extraction	51.7241	89.6292	0.001	1.36E+04	116.6779	2.49E+03
	Without Pattern	55.1724	9.37E+01	1.00E-03	1.42E+04	1.19E+02	2.58E+03

	Extraction						
S&P500_INDEX_closingPrice	With Pattern Extraction	44.8276	31.3878	5.26E-04	1.81E+03	42.556	330.6442
	Without Pattern Extraction	44.8276	3.15E+01	5.26E-04	1.82E+03	4.27E+01	3.33E+02

Prediction of day 5	Method	DS	MAE	MAPE	MSE	RMSE	NMSE
ALL_ORDINARIES_closingPrice	With Pattern Extraction	65.5172	234.5111	4.88E-04	8.42E+04	290.1763	1.54E+04
	Without Pattern Extraction	58.6207	232.3687	4.88E-04	8.31E+04	288.3121	1.52E+04
HangSeng_closingPrice	With Pattern Extraction	51.7241	1.61E+03	8.10E-04	4.36E+06	2.09E+03	7.95E+05
	Without Pattern Extraction	51.7241	1.63E+03	8.10E-04	4.43E+06	2.11E+03	8.09E+05
Nikkei225_closingPrice	With Pattern Extraction	48.2759	695.3529	0.0012	6.44E+05	802.5519	1.18E+05
	Without Pattern Extraction	44.8276	6.57E+02	1.20E-03	6.18E+05	7.86E+02	1.13E+05
NZ50_closingPrice	With Pattern Extraction	58.6207	101.7084	0.0011	1.85E+04	136.0455	3.38E+03
	Without Pattern	62.069	1.10E+02	1.10E-03	2.00E+04	1.41E+02	3.65E+03

	Extraction						
S&P500_INDEX_closingPrice	With Pattern Extraction	51.7241	33.1322	2.25E-04	2.09E+03	45.662	380.6701
	Without Pattern Extraction	51.7241	3.28E+01	2.25E-04	2.07E+03	4.55E+01	3.77E+02

Table 3 Experiment Results Comparison between the proposed method and the original method

Table 2 shows a number of trends. The major improvement caused by the geometric pattern extraction method is in the DS metric. Pattern extraction performed better in a large number of cases and was outperformed by SVR without pattern recognition in only a very few cases. Pattern extraction with SVR also performed better with respect to the MAE and RMSE metrics, although the difference between the two methods was not as large as with the DS criteria and was not as large as with the DS metric. There was no substantial difference between the two methods on the other three error function metrics that we tracked.

5.3 Summary

Observed from the above results, in our research, the prediction with the proposed geometric correlation extraction method does not appear to be predominant in performance as compared with the prediction without the proposed geometric correlation extraction method, even though for some accuracy metrics, namely the DS, MAS and RMSE, the SVR prediction with the proposed geometric correlation extraction method performance.

The lower than expected superiority of the geometric pattern extraction method may have been caused by the noise occurring in the correlations extracted and the over-fitting occurred during the data training process. More research is required to clarify this problem.

A possible method is to introduce the weighted Pearson's correlation method into the proposed method. This research suggests the weighted Pearson's correlation to cooperate with the geometric correlation extraction method in chapter 4, for bringing correlations to macroscopic factors in and filtering the noise in correlation knowledge. However, in the experiment section, weighted Person's correlation is not involved for some reasons. The corrections to the macroscopic factors may help to improve the accuracy (Song, 2010).

Chapter 6: Conclusion and Future Work

This chapter concludes the entire thesis, and the limitations of our method have been discussed as well. In order to overcome the limitations, some future work has been proposed.

6.1 Summary

It is difficult to predict the stock market movements as the data associated with stocks is the time series data characterized in by, noise, uncertainty, etc., and more particularly, the stock market is significantly influenced by the factors from both the microeconomic and macroeconomic environments, wherein the macroscopic factors are unpredictable somewhat. A number of automated computer-aided machine learning methods been developed for use of data process in such a chaotic environment.

The literatures show the success of the machine learning methods in stock analysis. However, the literatures also show that, from the historical data of a stock observed simply extracting some rules that the stock follows, and thereby forecasting the stock movements within a certain future time period, is rather difficult and exhibits lower accuracy. However, while studying a stock with its graphical representation (zigzag curve), the researcher found that the similarities of the stock movements are often found. Thus, a geometric method is developed to extract the movement similarity from the past data. With this geometric approach, an arc is modeled to approximate graphically the zigzag curve of a specific time length e.g. 25 days (in order words it is about using continuous function to represent discrete data approximately), thus there are in total n (n1-n25, n2-n26…) arcs drawn within the whole time range of the historical data (while an arc, as a part of a circle, is defined, the center and radius of the circle are determined therewith), and then SVMs (Support Vector Machines) is used to extract the correlation knowledge based on these circles, therefore to predict the movement of the stock (market trends).

As compared to the ordinary correlation extraction methods, the key improvement of the correlation extraction method proposed is the graphical trend similarity approximation, using an arc to represent the trend of a portion of the whole time series; it is called geometric correlation extraction method in this research. Within the conventional correlation extraction methods, the evaluation of the trend similarity between two time series is carried out by calculate the distance of the two time series to the time series observed. However, within the correlation extraction method proposed in this research, the evaluation of the

trend similarities is used to calculate the distance of the two arcs the two time series to the arc graphically approximating the time series observed. This means, the new method extracts correlation from general variation trends, avoids extra cting correlation form the time series directly, the mismatch problem is thus solved.

The comparison of the performance of the predictions measured by the performance metrics shows that the proposed pattern extraction method cannot meet the expectations set in the research, the unsatisfactory performance may be caused by the over-fitting problem, or noise in the extraction knowledge.

6.2 Limitations

The comparison of the performance of the predictions does not fully by the performance metrics shows that the proposed pattern extraction method cannot meet the expectations set in the research; the unsatisfactory performance may be caused by the over-fitting problem, noise in the extraction knowledge and other limitations. Those limitations have been discussed as below:

1. The stock prices are randomly distributed, which is a non-linear problem. For now, there still is not a perfect method to solve the problem with current techniques.
2. Usually, a non-linear problem is often solved by transferring it to a linear space. However, the conversion process involves too much approximation which impacts the performance of our method.
3. As known, the stock price is influenced by many factors, such as world economic status, Special Events (i.e, natural disaster), Investor Confidence and so forth. In our project, some factors, such as economic status, have been considered, but some factors are missing, like events and investor confidence.

6.3 Future work

In order to overcome the limitations discussed above, we have proposed some future work as follows:

1. To solve a non-linear problem, we often transfer it into a linear space. However, our method roughly converts the entire problem into a linear model. This impacts the performance of the method. Thus, the next step work could be further improving the performance and solve the problem. We could divide the entire problem into several small parts. Each part can be regarded as a linear system. For example, for an arc, a very small part of arc can be regarded as a straight line.

2. In our research, Marco economic has been considered which is often presented in the Indices of all stocks. However, there are two other factors have not been considers, investor confidence and events. In future, we could involve more considerations in these two factors. For investor confidence, we could conduct some qualitative research, such as surveys or questionnaires, to investigate. For consideration on special events, we could conduct event studies (MacKinlay, 1997). An Event study is often regarded as a statistical method for assessing the impact of an event.

Further research is required to clarify why the proposed pattern extraction method is unsatisfactory, and thereby to further improve the proposed pattern extraction method. A possible method is to introduce the weighted Pearson's correlation method into the proposed method. The combination of the pattern extraction method and weighted Pearson's correlation is expected to retain the current advantages, e.g., avoiding mismatch in correlation extraction, while have the advantages from weighted Pearson's correlation, e.g., filtering out the noise.

References

Abraham, A. et al., (2001), Hybrid Intelligent Systems for Stock Market Analysis, retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=6389D72358920E9F0E39A058C61A1DF7?doi=10.1.1.28.3141&rep=rep1&type=pdf.

Bantouna, A., Tsagkaris, K., & Demestichas, P., (2010), Self-organizing maps for improving the channel estimation and predictive modeling phase of cognitive radio systems, Proceedings of the 20th international conference on Artificial neural networks, Pages 382-391.

Chang, A.K., (2007), A study of grey theory on improving the investment performance of technical analysis index—an example of morgan stanley taiwan index' component stocks, GSIS 2007, pages 1422 - 1425.

Dodd, D., (2009), An Introduction to Fundamental Analysis and the US Economy, retrieved from: http://www.informedtrades.com/13036-introduction-fundamental-analysis-us-economy.html.

Glass, M., (2008), Stock Market - How to Use Fundamental Analysis to Make Trading Decisions, retrieved from: http://ezinearticles.com/?Stock-Market---How-to-Use-Fundamental-Analysis-to-Make-Trading-Decisions&id=1213032.

Guangyi, C., & Gregory, D. (2005). Auto-Correlation Wavelet Support Vector Machine and Its Applications to Regression. Paper presented at the Proceedings of the 2nd Canadian conference on Computer and Robot Vision.

Goetzmann, W. N., (1997), The Dow Theory: William Peter Hamilton's Track Record Re-Considered, retrieved from: http://viking.som.yale.edu/will/dow/dowpaper.htm.

Han, M., (2007), Multivariate Time Series Correlation Extract and Prediction Based on Cluster, Control Conference in 2007, pages: 187 – 191.

Hong, W., (2009), Features extraction and correlation analysis of stock index, Intelligent Control and Automation (WCICA), 2010 8th World.

Hopfield, J. (1982), Neural networks and physical systems with emergent collective computational abilities. Academic Science, Vol. 79, pp. 2554-2558.

Huang, C.-J., Yang, D.-X., & Chuang, Y.-T. (2007). Application of wrapper approach and composite classifier to the stock trend prediction Elsevier Ltd.

Huang, W., Nakamori, Y., & Wang, S.-Y. (2004). Forecasting stock market movement direction with support vector machine. Elsevier Ltd.

Hui, X., & et al., (2004), Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs, KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 334-343.

Jung-Hua, W., & Jia-Yann, L., (1996), Stock market trend prediction using ARIMA-based neural networks, Neural Networks, 1996, IEEE International Conference on Date of Conference: 3-6 Jun 1996, Volume: 4.

Kamruzzaman, J., Sarker, R.A, & Ahmad,I., (2003),. SVM Based Models for Predicting Foreign Currency Exchange Rates, Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), Pages 557-560.

Kim, H. J., Lee, Y. K., Kahng, B. N., & Kim, I. M. (2002). Weighted scale-free network in financial correlation. Journal of the Physical Society of Japan, 71(9), 2133-2136.

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. Elsevier B.V.

Kolarik, T., & Rudorfer, G. (1994). Time series forecasting using neural networks. *SIGAPL APL Quote Quad, 25*(1), 86-94. doi:10.1145/190468.190290

MacKinlay, C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature, 35*(1), 13-39. doi:citeulike-article-id:1557290

Neely, C.J (1998). "Technical Analysis and the Profitability of US Foreign Exchange Intervention". Federal Reserve Bank of St. Louis Review 80 (4): 3–17. Retrieved 2008-03-29

Nison, S., (1991). Japanese Candlestick Charting Techniques. pp. 15–18. ISBN 0-13-931650-7.

Pan, H. P. (2003), A Joint Review of Technical and Quantitative Analysis of Financial Markets Towards A Unified Science of Intelligent Finance, proceedings of 2003 Hawaii International Conference on Statistics and Related Fields.

Pai, P.-F., & Wei, W.-R. (2007). Predicting movement directions of stock index futures by support vector models with data preprocessing. Paper presented at the Industrial Engineering and Engineering Management, 2007 IEEE International Conference.

Rodgers, J., & Nicewander, A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician, 42*(1), 59-66. doi:citeulike-article-id:361042

Shin, K. S., Lee, T. S., & Lim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 127-135.

Song, Y., & Zhang, L., 2008, A Non-parametric Approach to Pair-Wise Dynamic Topic Correlation Detection, Proceedings of Eighth IEEE International Conference on data mining in 2008, Pages 1031 – 1036.

Song, L., (2010), Informative Correlation Extraction from and for Forex Market Analysis, AUT master thesis.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2008). Measuring and testing dependence by correlation of distances. Annals of Statistics, 35(6), 2769-2794. Retrieved from http://arxiv.org/abs/0803.4101

Székely, G., & Rizzo, M. (2009). Brownian distance covariance. *The Annals of Applied Statistics, 3*(4), 1236-1265. doi:citeulike-article-id:10257468

Tabbane, N., Tabbane, S., & Mehaoua, A., (2004), Autoregressive, moving average and mixed autoregressive-moving average processes for forecasting QoS in ad hoc networks for real-time service support, VTC 2004-Spring, 2004 IEEE 59th.

Tay, F. E. H., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. Elsevier Science.

Vapnik, V. (1998). Statistical Learning Theory. New York: Wiley-Interscience.

William, H., et al., (2007), Support Vector Machines, Numerical Recipes: The Art of Scientific Computing (3rd ed.), Cambridge University Press, New York.

Yung-Keun, K., Sung-Soon, C., & Byung-Ro, M. (2005). Stock prediction based on financial correlation. Paper presented at the Proceedings of the 2005 conference on Genetic and evolutionary computation, Washington DC, USA.

Yoo, P. D., Kim, M. H., & Jan, T. (2005). Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation. Paper presented at the Computational Intelligence for Modeling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference.

Yung-Keun, K., Sung-Soon, C., & Byung-Ro, M. (2005). Stock prediction based on financial correlation. Paper presented at the Proceedings of the 2005 conference on Genetic and evolutionary computation, Washington DC, USA.

Zhang, P., (2009), Extraction and Analysis of Shanghai Stock Index's Chaotic Parameter Information Technology and Applications, IFITA '09 Conference, Volume 2, pages 307 – 310.