Detecting Self-Correlation of Nonlinear, Lognormal, Time-Series Data via DBSCAN Clustering Method, Using Stock Price Data as Example

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of the Ohio State University

By

Shiyin HUO, B.E.

Graduate Program in Computer Science and Engineering

The Ohio State University

2011

Master's Examination Committee:

Dr. Rajiv Ramnath, Advisor

Dr. Jay Ramanathan

Abstract

In our modern world, the ability to predict the trend of nonlinear, irregularly shaped data is of great importance. This analysis can be used in many industries, such as in the prediction of stock prices, predicting a certain region's future electrical usage to make the grid more efficient and so on. This study is focused primarily on predicting rising and falling after a smooth period when the time series data shows a "turning point". The nonlinear, irregularly shaped time series data which we study in this thesis always show consistent and repeated rising and falling pattern over time. Thus we can predict the future rising and falling via studying the past data. This requires analysis of the self-correlation between the past time-series data.

In the last twenty years, many methods have attempted to predict this very aspect. Although some of these methods have shown good predictive accuracy, the high computational cost and low efficiency made them impractical outside the academic scope.

In this thesis, I have proposed an efficient method for analyzing self-correlation of a nonlinear, irregularly shaped time-series data using stock data as an example, which will select stocks and perform automatic trading. Rather than attempting to predict data future rising and falling at every moment, this method only runs the prediction function when some "turning point" becomes apparent. The DBSCAN clustering method is used to obtain self-correlation between current and previous segments of time series data. By calculating the average expected rate of return for similar segments, we derive the

predicted parameters. To select the optimal stock for trading, we need to simulate how the stock will perform in the future with the predicted parameters. At last, we select the stock whose future performance is best.

The contributions of this thesis are as follows:

1 This thesis proposes a method to automatically locate data peaks and valleys.

2 This thesis also proposes a method to analyze self-correlation of nonlinear, irregularly shaped time series data with highly reduced computational cost.

3 Finally, optimal stock selection is achieved by analyzing the stock's predicted expected rate of return and its volatility.

Dedication

This document is dedicated to

my father Guicheng Huo and my mother Lijie Yang.

Acknowledgments

I want to express my deepest gratitude to my advisor Dr. Rajiv Ramnath and Dr. Jay Ramanathan for their support and friendship during my study and research.

Also, I would like to thank all my friends from CETI. Your help and friendship will never be forgotten.

Lastly, I would like to thank my parents for their unwavering support, not only in my graduate studies but in all of my endeavors.

Vita

July 2009.....B.E. Automation, Tsinghua University

September 2011 .....M.A. Economics, Ohio State University

Fields of Study

Major Field: Computer Science and Engineering

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Thesis Organization

This thesis consists of six chapters. A chapter summary is as follows:

Chapter 1 presents the overall organization of the thesis and fully describes the issues from both the economic and computer science perspective. At last, the main contributions of this thesis are discussed.

Chapter 2 begins with some background information about the stock market and stock trading, and later focuses exclusively on the China stock market from where this study's data was collected. Secondly, the reason for using stock price data as the example of nonlinear, irregular in shape data is explained. Thirdly, the reason why nonlinear, irregularly shaped data can be predicted is discussed. At last, some background information is presented about the clustering method.

Chapter 3 primarily discusses the previous studies related to this topic.

Chapter 4 gives a detailed description of the structure and the algorithms used in the system.

Chapter 5 shows the trading and the prediction result.

Chapter 6 concludes the thesis with suggestions for future research.

1.2 Problem Description

1.2.1 Overall Description of Problem

The nonlinear, irregularly-shaped time series data that is the focus of this thesis can be illustrated by a specific model that consists of two parameters: the growth rate, which is non-stochastic; and the volatility, which represents the stochastic part of the model.

Future growth rate can be predicted by analyzing the previous data. However, the exact amount of future data cannot be predicted because of its stochastic nature. Thus we are faced with two major problems. One is to predict the future growth rate and volatility of the data; the other is to simulate future data by running the model according to the parameters, which have already been prescribed.

These two problems will be explored in detail in section 1.2.3 and 1.2.4

1.2.2 Automatic Detection of Data Peaks and Valleys

Predicting data's future growth rate at every moment requires a massive computational cost, which unfortunately will not correspond to massive trading opportunities and will result in output of a huge amount of irrelevant information.

The solution is to find some peaks and valleys of data which indicate potential trading opportunities. The prediction function will only be applied when "peaks or valleys" appear.

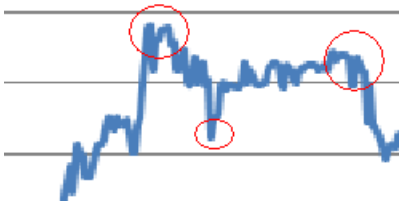

Figure 1 Peaks and Valleys

For example, the two red circles above indicate “Peaks” of data, while the circle below indicates the “Valley” of data. People are more interested about whether the future data will rise or fall after these “Peaks and Valleys”.

In addition, the trade-off between computational cost and trading opportunity loss should be more balanced. That is, too many “peaks or valleys” will slow down the system; while too few “peaks or valleys” will miss potential profiting opportunities. Therefore, the number of “peaks or valleys” being detected should be maintained at some reasonable level.

1.2.3 Detecting Self-Correlation of Nonlinear, Irregularly Shaped, Time Series Data

We can predict the future growth rate of nonlinear, irregularly shaped time series data by analyzing its self-correlation.

If we assume that stock prices maintain consistent behavior, future growth rate can be predicted based on the similarity between current data and previous data. To enhance the predictive accuracy, clustering previous data into categories is necessary. The data containing relatively large self-correlation should be placed into the same cluster. The

clustering method should be chosen carefully in order to obtain optimal clustering accuracy.

The next step is to compare the current data with previous clusters. We can predict that the future rate of return will be the same due to the cluster's average rate of return whose Euclidean distance with current data is smallest.

1.2.4 Simulation

Although we can predict the growth rate of data, we still don't know how the future time series data will perform. Since the future time series data has stochastic character, a predicted value only is not a strong enough basis on which to select stocks.

The data with low growth rate but high volatility has possibility to reach some high value in the future. Thus the return of a stock with high volatility but a low rate of increase may exceed the return of a stock with low volatility but a high rate of increase for the same period.

If the predicted rate of return and volatility are used as input parameters, the simulation functions can be run to derive predictions of future performance.

In addition, other than the self-correlation in one stock, there are correlations between different stocks. For example, when the petroleum price rises, people will choose to fly less. That is, when the stock prices of petroleum rise, the stock prices of airline companies will fall simultaneously. These correlations between stocks should be preserved during the simulation.

1.2.5 Stock Selecting Problem

How to select a high performance stock has been highly sought after for many decades. Many theories and methods have been proposed just for selecting the optimal stock for different investment purposes. Basically there are two main approaches to the selection of stocks: the qualitative method and the quantitative method. These approaches will be discussed in detail in Chapter 2.

For now, it is sufficient to say that the quantitative selection method depends on using statistical analysis of the stock's behavior to predict its future trend. This requires exploring the self-correlation between current and previous stock price time series data in order to obtain similarities between these segments of data. At last, we can predict a stock's future rising or falling under the assumption that the stock maintains consistent behavior.

1.2.6 Trading Policy

Having obtained the predictions, the next step is to determine a trading policy. However, sometimes a stock simply will not follow its predicted course; Sometimes its behavior will be completely unpredictable.

An example of this would be if an investor buys many shares of a stock based on its predicted rise, however, the stock price begins to fall. When this occurs, two scenarios are possible. First, after a transitory decrease period, the price will increase and will be ultimately higher than the purchase price. Secondly, the stock price will perform under

the purchase price for a long period of time. Therefore, the investor should determine a threshold of whether to close trading or continue to wait.

On the contrary, if the stock price starts to rise after purchase, there are also two possibilities. It may keep rising for some time. However, nobody knows when it will reach its peak. A rushed sale may lose potential profit. On the other hand, waiting to sell may not only result in lost profit, it may even cause the price to fall below the purchase price. So a threshold is indeed necessary.

1.3 Thesis Contribution

1.3.1 Purpose a new method for detecting self-correlation of nonlinear, irregularly shaped, time-series data

From previous studies, it is understood that the time series data which we study in this thesis exhibits at least the following characteristics:

1 Time series data is highly nonlinear.

2 The stock time-series which have similar tendencies don’t form a regular shaped cluster. However, they gather in higher-density irregular shaped clusters.

These three characteristics make the density-based clustering method a suitable one.

However, it is still unknown how many clusters are there. Arbitrarily assigning the number of clusters will result in lower predictive accuracy. A clustering method is needed which does not require initially assigning the number of clusters. Thus, the “density-based spatial clustering of applications with noise,” (DBSCAN), is ideal for this research.

By employing the DBSCAN method, a fixed length of stock price data can be compared with past clusters, which can yield the predicted growth rate.

1.3.2 Determining an Effective Way to Detect data Peaks and Valleys

Attempting to predict a stock's rising and falling at every moment will result in an unnecessarily high computing cost. Otherwise, predicting the behavior at arbitrary times may lose profitable trading potential. Deciding when to activate the prediction function will balance the trade-off between trading profitability and computing efficiency.

This thesis proposes an effective way of detecting data "peaks and valleys". We will activate the prediction function when these "peaks and valleys" are shown.

By this procedure, our system will reduce the computing cost substantially and create promising earning potential.

1.3.3 Synergy of DBSCAN and the Monte Carlo Simulation

Previous research of mining the stock market stops at "predicting" a stock's price or tendency. Some of these predictive methods do have excellent accuracy. However, they have a major drawback. The movement of stock prices is a highly stochastic process, but these methods all give a static prediction price. In other words, they ignore the impact of volatility on stock prices. Take for example two types of stocks: stock A has a higher than expected growth rate but lower volatility; stock B has a smaller than expected growth rate but higher volatility. Although in the long run stock A may outperform stock

B, in the short run, investing in stock B may earn more than investing in stock A because of its larger possibility of fluctuation.

This will mislead investors, sometimes even causing them to select the wrong stock. So this thesis uses a clustering method to estimate parameters. The Monte Carlo method uses these parameters to simulate future stock prices. The system then selects the stock with the best simulation results.

Chapter 2: Theoretical and Related Background

2.1 Data Characteristics

2.1.1 The Markov Property of Data

A stochastic process is a time sequence of values that are subject to some stochastic distribution at each time point. For example, the number of customers arriving at a shop every minute conforms to a Poisson process. The Markov process is also a stochastic process. It assumes that the distribution of future value is only determined by the current value. That is, past history has no direct relevance to what will happen in the future. The mathematical condition of Markov property [1] is:

$$\begin{aligned} &\Pr[X(t) = x(t) \mid X(s) = x(s)] \\ &= \Pr[X(t) = x(t) \mid X(s) = x(s), X(p_1) = x(p_1), X(p_2) = x(p_2), ...] \end{aligned} \tag{2.11}$$

The relationship of time is: $\cdots\cdots < p_2 < p_1 < s < t$

Stock prices are usually assumed to have this Markov property [2]. Since, in a short time period, there is seldom newly released information that could cause investors to change their minds about which stocks to buy. Thus the Markov process is a good illustration of a stock price's short-term fluctuations. In addition, the Markov process supports the weak form of "Efficient Market Hypothesis (EMH)," which will be discussed in next section. Because the Markov process assumption of stock prices states

that past performance has nothing to with future performances, this is consistent with EMH's statement of current stock price reflect all the related information of such stock.

However, the Markov process assumption is not perfect with regard to stock fluctuations because in the long term, fluctuations of stock price are driven by fundamental changes in the world such as interest rate, earning ability of the underlying firm, and macro economy conditions. It must be noted that the Efficient Market Hypothesis is still very controversial.

2.1.2 Lognormal Distribution of Data

Basically, stock prices are driven by two forces. In the long term, stock prices are driven by fundamental changes, while short-term uncertainty is caused by speculation by investors.

If we assume that the increasing rate driven by fundamental changes is μ, then if there is no uncertainty, we should have the increment of stock price ΔS in time interval Δt as follows:

$$\Delta S = \mu S \Delta t \tag{2.1}$$

Take limit, as $\Delta t \to 0$, we have:

$$dS = \mu S dt \tag{2.2}$$

In reality, of course, there is always uncertainty. We add uncertainty to our model as an increment from the Wiener process, which is defined as follows.

A stochastic process z is called the Wiener process [3], if it satisfies two properties:

A: During a small time interval Δt, the change is

$$\Delta z = \varepsilon\sqrt{\Delta t} \tag{2.3}$$

ε is a standard normal distribution

B: The values of two Δz are independent and from any two different time intervals.

After modification by adding the Wiener process variable, we can derive the famous discrete-time version stock price behavior model [4]:

$$\Delta S = \mu S\Delta t + \sigma S\varepsilon\sqrt{\Delta t} \tag{2.4}$$

or

$$\frac{\Delta S}{S} = \mu\Delta t + \sigma\varepsilon\sqrt{\Delta t} \tag{2.5}$$

Where μ is the underlying stock's expected rate of return per unit of time, σ is that stock's volatility.

By ITO's Lemma (Appendix A), we can further derive the lognormal distribution of the stock price [5].

Define

μ : Expected rate of return

σ : Volatility of the stock

S: Stock price

T: Time interval

$\phi(m,v)$: A normal distribution with mean m and variance v

The lognormal distribution is:

$$InS_T \sim \phi(InS_0 + (\mu - \frac{\sigma^2}{2})T, \sigma^2 T) \tag{2.6}$$

The figure of this distribution is as follows:

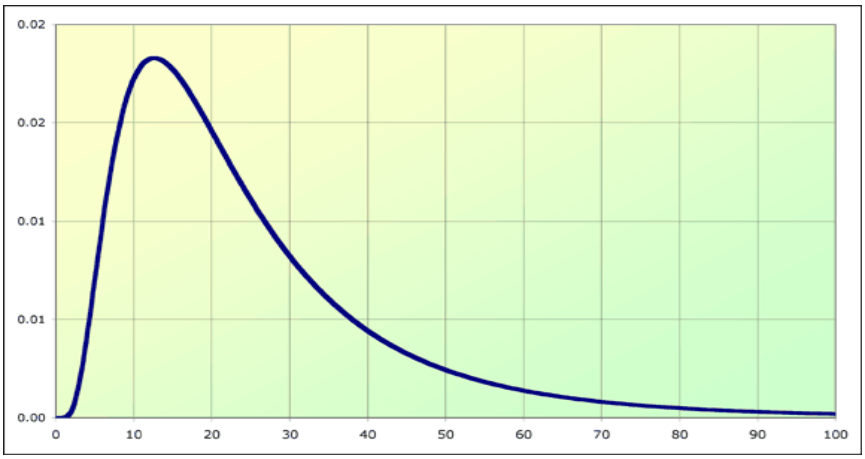

Figure 2 Lognormal Distribution

There is one thing that must be mentioned at this point. Actually, it is the percentage change of stock prices which follows the log-normal distribution, not the stock price itself. We say “log-normal distribution of stock price” just for convenience.

2.1.3 Nonlinear Property

In reality, “a nonlinear signal” is equal to “a nonlinear system.” A time series data has nonlinear property means that the data can be characterized by a nonlinear system. For example, if the stock price is determined by the underlying firm’s profit, then the firm’s profit is the input of the system and the stock price is the output. If the system is

linear, then an increase of profit will lead to the stock price increases by the same proportion.

Let's now compare linear and nonlinear systems. Generally, there are three components of a system: input, operations and output. Typical linear systems have three necessary conditions: superposition, homogeneity and a unique equilibrium.

Firstly, linear systems conform to the superposition condition. For two independent inputs $f_1(t)$ and $f_2(t)$, the system has two outputs $y_1(t)$ and $y_2(t)$ accordingly. If we take the superposition signal as input, for a linear system, we should also have the superposition output as follows:

$$f_1(t) + f_2(t) \xrightarrow{Linear\ System} y_1(t) + y_2(t) \tag{2.7}$$

This means that there is no interaction between independent signals when passing through a linear system. However, for nonlinear systems, there are in fact interactions between signals. So the nonlinear system may look like this:

$$f_1(t) + f_2(t) \xrightarrow{Nonlinear\ System} y_1(t) \cdot y_2(t) \tag{2.8}$$

Secondly, linear systems are homogenous. A system which is homogenous of degree n should satisfy the following condition:

$$f(\alpha v) = \alpha^n f(v) \tag{2.9}$$

Intuitively, homogeneity means if the input was multiplied by a certain number, then the output will also be multiplied by that same number. However, nonlinear systems do not have this property.

Thirdly, linear systems have a unique equilibrium point.

$$f(v) = v \tag{2.10}$$

The nonlinear systems do not have this property.

For the stock price data, “equilibrium point” means holding the input (underlying firm’s profit, past stock price and so on) as constant, what the stock price will be.

However, there are many empirical studies that support the result that the system of stock prices is a nonlinear system. Lye and Martin [6] give several economic reasons about why stock price models are nonlinear:

(1) Investors will jointly consider the mean (expected return) and the variance (risk) of the stock while deciding invest in a stock. Since the joint distribution of mean and variance is not linear, the stock price system will not be linear either.

(2) Nonlinear models can better capture investor’s behavior of outliers and overreactions. [7]

(3) Investors have nonlinear response to newly released information. [8]

(4) The “reversibility” characteristic can be derived from a linear system, since, to cancel an impact, a stimulus is needed which is just the opposite of the original input. However, investors won’t exhibit such behavior.

(5) In the real investment world, there are many equilibria; while the linear system only has only a single equilibrium point.

So in this thesis, the stock price model is interpreted as nonlinear to better capture investors' behavior.

2.1.4 Irregular in shape

We are not interested in a single value of the price of a stock. We are also not interested in a very long period of stock prices either. What we are really interested is a certain length of stock price data from which we can extract some consistent behavior. For example, we can divide the total stock price data into vectors that are of two minute length (There are two points in each vector, since one-minute stock price data is used in this thesis.).

Date	Time	Price
2011-1-4	9:31	12.7
2011-1-4	9:32	12.72
2011-1-4	9:33	12.7
2011-1-4	9:34	12.65
2011-1-4	9:35	12.64
2011-1-4	9:36	12.66
2011-1-4	9:37	12.66
2011-1-4	9:38	12.66
2011-1-4	9:39	12.62
2011-1-4	9:40	12.6

Table 1 10-minute data example

For example, Table1 shows a ten-minute data example. We can divide it into five two-minute vectors: (12.7, 12.72), (12.7, 12.65), (12.64, 12.66), (12.66, 12.66), (12.66, 12.62).

If we plot the “two minute” data on a two dimensional space, then we get a point. The points which are close in Euclidean distance can formulate a cluster. A cluster of these points can gradually take a particular shape. For many systems, the shape of clusters is spherical as in Figure 3.

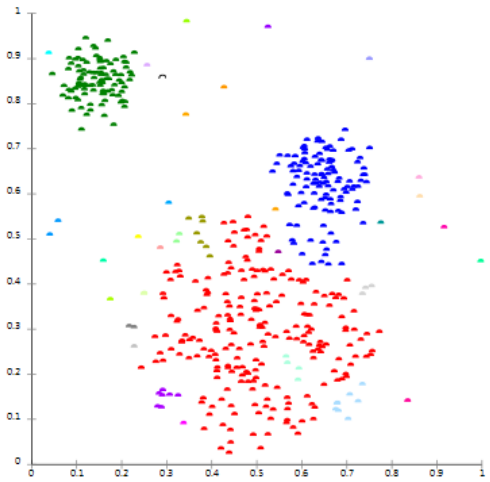

Figure 3 Spherical Shaped Cluster

However, there is no evidence shows that the cluster of stock price data will have a regular shape. The clusters may appear as follows:

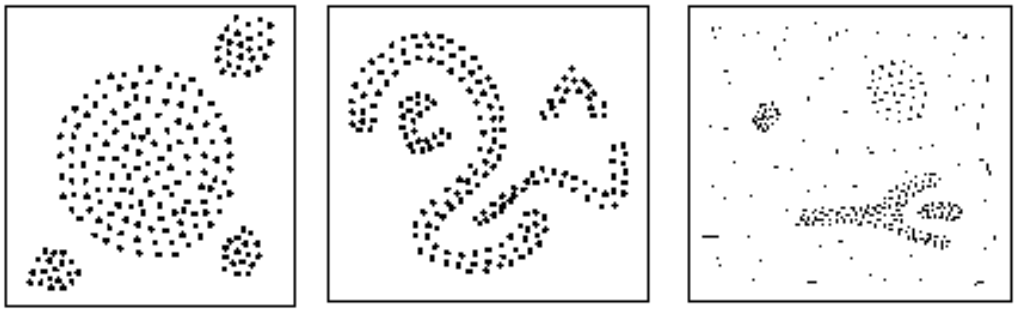

Figure 4 Irregular Shaped Clusters

The clustering of data which has irregularly shaped clusters is relatively easy. The k-NN clustering method will do. However the k-NN method has low clustering accuracy on irregular shaped clusters.

Since the clusters' shape of stock price data is unknown and the number of clusters is hard to be predetermined, we need an appropriate and more complex clustering method.

2.2 Efficient Market Hypothesis and Behavioral Finance

Finding significant patterns in past stock data and developing active trading strategies which have returns higher than the index are always the dominant challenge in the finance world [9].

To determine the intrinsic value of a stock is impossible. Consequently, testing whether the stock price reflects the true value is also impossible [10]. Thus, most of these types of tests concentrate on the performance of trading strategies. Interestingly, there are

some effective trading strategies and famous investors (George Soros, Warren Buffett) which have yielded superior returns.

However, there is a trade-off between risk and returns. The stock has a larger risk than the bond. To attract investors to buy, the stock must have a greater return than the bond. The difference of returns between risky asset (stock etc.) and risk free assets (bond etc.) are called “risk premium” [11]. No study can determine whether these “superior returns” made by investors are from pattern finding or just a reflection of risk premium.

The Efficient Market Hypothesis holds the opinion that prediction of the market is useless while the Behavioral Finance theory claims that prediction the market can lead to excess profit. Both the Efficient Market Hypothesis and Behavioral Finance theories have some empirical studies to support them and their own advocates.

2.2.1 Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) [12] supposes that all the information which is useful for predicting the future stock price has already been reflected in the current stock price. The stock price will only adjust for newly released information. When there is no new information, the stock price should represent a “random walk,” which means the future direction is unpredictable and the best expectation of tomorrow’s stock price is today’s price. However, “unpredictable” is not the same as “irrational.” It just means the current price reveals the intrinsic value of that stock.

There are three versions of EMH.

The weak form of Efficient Market Hypothesis states that the current stock price reflects all the market trading data including trading volume, stock price time series data, price-to-earning ratio and so on.

The semi strong form of EMH asserts that besides market trading data, all publicly accessible information about the associated firm has already been reflected in its current stock price. This means that the information revealed in its financial statement and other information from public news are thoroughly revealed by the stock price. Therefore, further analysis of the firm from public information is a waste of time.

The strong form of EMH contends that, besides market trading data and public information, insider information about the associated firm has also been reflected by the current stock price.

Most proponents of the Efficient Market Hypothesis believe in the semi-strong form of the hypothesis.

The conclusion derived from EMH is that in order to make money in the stock market, you have to either get information faster than others or analyze data more thoroughly than the other competitors.

Most importantly, EMH asserts data mining from past stock time-series is fruitless.

2.2.2 Behavioral Finance

The core assumption of Efficient Market Hypothesis is that investors price the stock exactly at its “true value,” which is of course, impossible in reality. In response to this critique, the Behavioral Finance theory [13] arose. This theory focuses on studying

investors' behavior when making decisions on risky investments. There are five main causes of "investor irrationality". [14]

The first one is that people process information incorrectly. Due to limitations of time and other resources, investors can only analyze a small sample of stock price data and use this limited analysis to understand the whole picture. Thus the representativeness of the sample affects whether investors receive correct information to begin with and if they can process it accurately or not. In addition, investors are not sophisticated enough to analyze the situation correctly. For example, an increase in interest rates can cause mortgages to rise; consequently, the housing market will be negatively affected. After this, the number of new houses being built should decrease, which will cause construction materials supplier's expected earnings to decrease. As a result, these suppliers' stock prices will fall. However, not all investors can comprehend these complicated relationships and make the correct decision.

Secondly, investor's are risk-averse. This means that normal investors require a risk premium on risky assets. For example, there are two choices: taking $10,000 as a sure thing or having a 50-50 probability of tripling this amount or none. Most investors will opt for the first choice even though the second one has a higher expected return. Though the investor will always make the optimal choice to maximize his or her own safety, this choice will not always be the one with the best expected return rate. Thus the current stock price may not reflect its intrinsic value.

Thirdly, the psychological factors involved in making investment choices have an influence on stock prices. Here I will introduce some common investing psychology.

When an investor makes a successful investment, he or she is likely to credit it to intelligence. However, when a bad investment is made, the investor will blame bad luck. Day by day, people tend to overestimate their own abilities. Thus people will not process the information irrationally; the stock price will contain investors' trading behavior.

In addition, recent earnings or losses can make investors more willing to take risks or remain conservative. This will affect investor's rationality.

Furthermore, Behavioral Finance theory hypothesizes that investors tend to be more regretful when an unconventional decision was proven to be a bad one. This phenomenon would lead to what is called "regret avoidance." Thus individuals are much more risk-averse when an unconventional choice is on the table.

However, Behavioral Finance theory is still new. Whether or not the investor's behavior affects stock prices is still a controversial topic. However, the statement that investors are not fully rational in decision making is well accepted in the world of finance.

2.2.3 Technical Analysis

If Behavioral Finance theory is proven true, then behavior of investors will have a huge affect on stock prices. Investor's trading behavior will be reflected in the trend of the stock price. On the other hand, the stock price's trend can ignite an individual's trading activity by affecting their prediction of future stock price direction. Thus some patterns will be formulated in both directions.

These assumptions are just the theoretical basis of technical analysis. This analysis focuses on finding predictable and recurring patterns in past stock price time-series. The

famous Technical Analysis theory was purposed by Guru Charles Dow [15]. Most of the current technical analysis models are variants of the “Dow theory.”

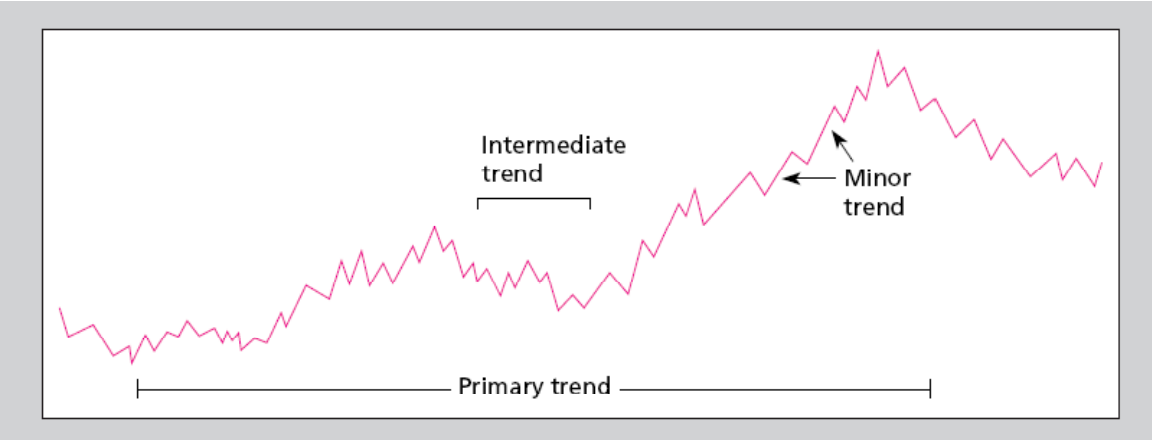

Figure 5 Dow theory trends

The Dow Theory decomposes the factor which will affect stock price into three forces:

1 The primary trend will affect the stock prices in the long-term, lasting from several months to several years.

2 Intermediate trends are stock price’s short-term deviations from the underlying trend line.

3 Minor trends are daily fluctuations caused by trading activity.

By analyzing these trends, investor can get a prediction of future stock price's direction.

2.2.4 Data Analysis by Computer

One of the most important conclusions derived from Behavioral Finance theory is that patterns exist in stock price data. Besides technical analysis, the data analysis method can be used in the computer science field to capture these patterns.

Recently, many methods such as data mining and clustering have been used for detecting useful patterns. These methods will be discussed in depth in the next chapter.

2.3 Cluster Analysis

Cluster analysis (clustering) is a data analysis method. The goal of clustering is to place similar objects into the same group. By clustering, we can retrieve useful information from irrelevant noise or uncover patterns of data. Take the following figure as an example; we can retrieve the data to its original group.

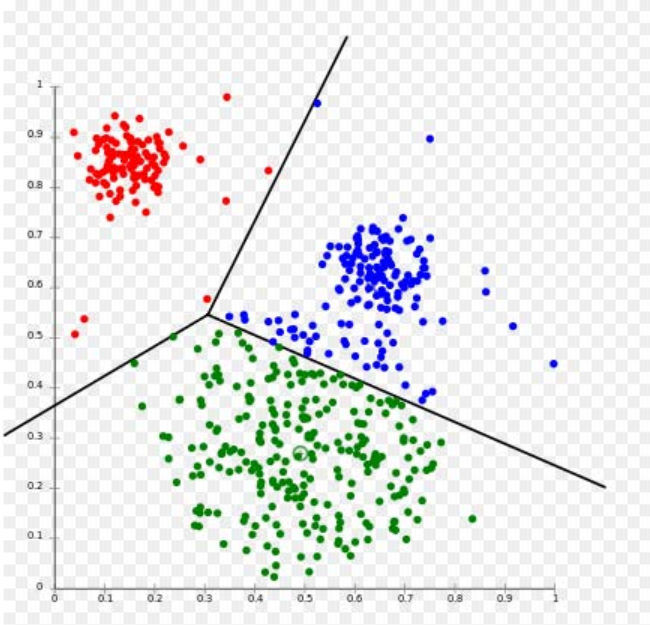

Figure 6 Clustering Example

The algorithms of clustering vary from type to type. Each problem must be carefully assigned an appropriate clustering algorithm. Table 2 gives some typical clustering models [16].

Model	Theoretical Background	Example
Connectivity Model	objects with close distance connectivity are put in same cluster	Hierarchical Clustering
Centroid Model	clusters are represented by their own id vectors	K-means Clustering
Distribution Model	clustering by statistic distributions	Expectation-Maximization Clustering
Density Model	put density-reachable objects into same cluster	DBSCAN
Subspace Model	add relevant attributes into clustering process	Biclustering

Table 2 Clustering Algorithms

In this thesis, we use apply density clustering models to stock time-series in order to find stable patterns for predicting future trend. We will discuss the algorithm in detail in Chapter 4.

2.4 Brief introduction of the stock market

2.4.1 The Stock Market

The stock market [17] is made up of the primary market and the secondary market. In the primary market, shares of stock are open for sale to the public for the first time. This is managed by the Initial Public Offerings, or IPOs. Companies sell out shares and raise money by offering public shares. Then these already-issued securities are traded between investors in the secondary market. Usually, what the average investor refers to as the “stock market” is actually the secondary market.

Based on the trading characteristics, we can further differentiate three types of markets: dealer market, brokered market and auction market. In any organized market, sellers can seek out interested buyers and vice versa.

The dealer market is organized by a group of dealers from which investors buy or sell stocks. The dealers purchase assets for their own accounts, and later sell them for a profit from their own inventory. Usually, the dealers are the top level traders in the market. They may enter the bid (buy price) and determine a price (sell price) at which they are willing to buy or sell the stocks, which are registered into the computer network and may update these quotes as desired.

In the brokered market, a financial agent or broker offers search services to buyers and sellers. A good example is the real estate market, where the prospective buyers are making such a large monetary investment that it is worthwhile for participants to pay brokers (real estate agents) to conduct the searches. Today, large block transactions are completed in the brokered market, in which vary large blocks of stock are brought or sold. When there is a trader who wants to sell a large block of stocks, the “block house” often is engaged to search directly for other large scale traders, rather than bring the trade directly to the market.

The most integrated market is the auction market, in which all traders converge at one place (Today, they can be connected electronically.) to buy and sell stocks. The reason for the popularity of the auction market is that the trading system matches buy and sell orders based on the “highest in price, earliest in time” policy. For example, when a

buy order is brought to market, only the sell order with the highest bid price can win the trade.

2.4.2 Stock Trading

Stock trading is completed by “match and execute” process of trading orders. Buyers enter the market with a buy order, while sellers have a sell order. Basically, there are two types of orders: market orders and price-contingent orders.

Market orders are orders which will be completed specifically at the current market prices. They are price takers. That is these orders take the market price as given.

The price-contingent orders are orders which specify price conditions on which to buy or sell securities.

2.5 China Stock Market

The Chinese stock market consists of three exchanges; the Shanghai Stock Exchange, the Shenzhen Stock Exchange and the Hong Kong Stock Exchange. The Hong Kong Stock Exchange is very different from the two mainland stock exchanges. Also the Shanghai Stock Exchange has a much larger total market value than the Shenzhen Stock Exchange. Therefore, this thesis will mainly focus on the Shanghai Stock Exchange. The following section will introduce the dominant features and concepts necessary to understand the largest stock exchange in China.

2.5.1 Auction Trading Policy

The Shanghai Stock Exchange uses auction trading, which can be conducted either as a call auction or a continuous auction. Whichever type of auction is used, execution is based on “price priority and time priority.” The price priority means a higher buy-price order has priority over a lower buy-price one, while lower sell-price orders have priority over higher sell-price options for a specific stock. For the orders with same price, the earlier placed orders have priority over later ones.

The process of one-time centralized matching of buy and sell orders during a specified period is called a call auction.

The execution price policy is determined by the following principles, during the call auction period.

(1)The price at which the greatest trading volume can be generated;

(2)The prices at which all buy orders are to be completed at a higher bid price..

(3)The prices at which all sell orders are allowed to be completed with a lower bid price.

(4)The price at which either all sell orders or all buy orders with identical prices can be completed at the same time.

If there is more than one price which satisfies these conditions, the execution price is determined as the price which can minimize the unexecuted volume. However, if there is more than one price that minimizes the unexecuted volume, we take the mean price as the execution price.

All the buy and sell orders are executed at an identical execution price during the call auction period. The remaining orders automatically enter the continuous auction.

Continuous auction refers to the process of continuous matching of buy and sell orders on a one-to-one basis. The execution price policy is determined by the following principles, during continuous auction period:

(1) The execution price is determined when the highest bid price matches the lowest offer price.

(2) If the currently available lowest offer price is lower than the bid price, the lowest offer price is accepted as the execution price.

(3) If the currently available highest bid price is higher than the offer price, the highest bid price is accepted as the execution price.

2.5.2 Trading Time

The Shanghai Stock Exchange opens for trading on regular business days. The trading hours are as follows:

9:15 – 9:25	call auction
9:30 – 11:30	continuous auction
11:30 – 13:00	rest
13:00 – 15:00	continuous auction

Table 3 China Stock Market Trading Hours

2.5.3 Order Type

In order to trade stocks, investors should initiate securities accounts and sign broker-client agreements with a brokerage company. Clients may place a limit order or market order as needed. An order placed by an investor should include the following items:

(1) The client's securities account number;

(2) The code of a particular security;

(3) Specify buy or sell order;

(4) The instructed quantity;

(5) The instructed price

(6) Any other information as required by the Exchange.

2.5.4 Special Trading Policy

There are two trading policies regarding timing. The "T+1" trading policy means the stocks bought today cannot be sold until the next business day. On the other hand, the "T+0" trading policy means an investor can sell and buy stocks without any limitation with regard to time.

Until 1995, China was under the "T+0" policy until it changed to the "T+1" policy, which it presently adheres to.

In addition, if any single stock rises or falls beyond 10% of its opening price in one day, then this stock's trading will be suspended for the rest of that day.

2.5.5 Market Statistics

Some information of the Shanghai Stock Market is shown in Table 4.

Total Market Capitalization(RMB100 million)	180905.08
Total Free-float Capitalization(RMB100 million)	147669.29
Total Turnover Volume(10,000 shares)	816546.56
Total Turnover Value(RMB 10,000)	9583435.63
Total Number of Trades(10,000 deals)	514.75
Turnover Rate	0.47
Average P/E Ratio	16.52

Table 4 Market Statistics of China's A Share, up to 7.21.2011

Chapter 3: Related Research

The related work on predicting stock prices and detecting self-correlation of nonlinear time-series data can be categorized into five groups. The remainder of this chapter will be devoted to exploring these categories in detail.

3.1 Decision Tree

Decision tree is an algorithm which is depicted as a tree-like graph. By comparing each decision and its possible outcomes, the decision tree outputs which option is the best course of action.

Corporate financial statements and other public information accessible to the public of any company are critical in the decision-making process. However, investors face two problems. First, the accuracy and efficiency of information are hard to determine. Secondly, processing mass amounts of information is an impossible task for the typical investor. Because of these issues, valuable trading opportunities may be lost.

To combat this problem, Jiang and Wang (2009) [19] proposed an improved Begging Decision Tree algorithm to predict the EPS (earnings per share) of one stock.

Rachlin, Last, Alberg and Kandel (2007) [20] came up with a C4.5 decision tree system the purpose of which is to predict stock price trends and make trading decisions based on the mining of web documents and data mining of a particular stock.

3.2 Clustering

Cluster analysis algorithms are mostly used to detect patterns and self-correlation in stock price data.

Stock price time series can be conceived of in two parts: one is the stable and inevitable trend caused by long-term factors; another is stochastic caused by trading and speculation noise. We can call the first part “motif” of time series. Since motif is consistent and shows high levels of self-correlation between different time-periods, it is critical to predicting future trends. Jiang, Li and Han (2009) [21] proposed a method of stock tendency prediction using ordinal comparison and k-NN clustering to uncover time series motifs.

Further, Huang, Jane and Chang (2007) proposed a new method [22] to be combined with Rough Set. This method forecasts the future trend, and then clusters the data by the k-means clustering algorithm. At last, the data was supplied to a rough Set to select the optimal stock.

Similarly, Wu, Denton and Elariss (2009) model the motifs of time series into several “core patterns” [23]. They proposed a method to decompose these patterns based on the density clustering algorithm.

3.3 Association Rules

Association rules are the relationships between variables in a database. In the stock market, it refers to the relationship between transactions. For example, a typical

association rule may be “a 15% increasing of stock XYZ will lead to a 5% decrease sequentially.” Recently, many researchers have attempted to mine association rules for the stock market.

In 2011, Voditel and Deshpande purposed a system for mining association rules and outputting investing recommendations. The system uses broad index and indexes of each sector as input. It detects the relationship between each stock and each sector’s index movement by association rules of mining algorithms along with fuzzy classification methods [24].

Similarly, Li, Xing and Huang (2010) proposed a system of mining association rules for similar stocks [25]. For example, the system could output association rules such as “If stock A’s price increases on day one, then stock B’s price has a 50% chance of falling in the next two days.

Romain Pierrot and Hongyan Liu (2008) purposed a method focusing on mining the relationship between the momentum of trading volume and stock volatility and a stock price’s future performance [26].

3.4 Neural Networks

Among the many computer science methods aimed at predicting stock prices, the Neural Network methods are the most widely used and have the best predictive accuracy.

Ikuo Matsuba devised a system using past stock as input, then predicting the future price, based on neural networks [27].

Then Jung-Hua Wang and Jia-Yann Leu (1996) proposed a modified system based on Matsuba's method that changed the input into stock price's second difference, and then predicts stock price by ARIMA-based Neural Networks [28].

In year 1999, Yu, Chen, Wang and Lai modified the original Neural Network algorithm by squares support vector machine to predict stock price's future trend [29].

FS and Goh (1991) used fuzzy sets to extend the Neutral Network method by analyzing and predicting stock price based on fuzzy probabilistic rules and Bollean data.

Nguyen, Omkar and Hayfron-Acquah (2006) further modified the Neural Networks. They proposed a system using Hierarchical Cerebellar Model Arithmetic Controller (HCMAC) neural network to predict future stock price [30].

Liao and Wang (2008) added stochastic factors into the original Neural Network model. They combine one stochastic time effect function and the forecasting model together, to get an improved predictive model known as the "stochastic time effective neural network".

The last but not the least, Zhou and Zhang (2010) propose a system to predict stock price's future trend using BP Neural Network. It takes stock price data and variables derived from it as input [31].

3.5 Textual Rules

The predictive method used for stock prices based on textual rules is a relatively new approach. Its core idea is using news, web information and other textual materials as

input, combined with multiple methods of analysis, to predict the future trend of stock prices.

Wuthrich, Cho and Leung (1998) proposed a stock price predicting system based on web messages [32]. Later, Fung, Yu and Lam (2003) devised a system analysis of the inter-relationship between stock price's movement and textual articles based on statistical significant [33]. The purpose was to announce the recommended stock after up-to-date news had been released.

Chapter 4: System Description

4.1 Overall Structure

4.1.1 Data Format and Stock Used

Stock data was employed as an example of non-linear, non-regular in shape time series data. The raw stock data used in this thesis is minute-to-minute, which consists of eight parts. A ten-minute example is as follows:

Date	Time	Start price	Highest price	Lowest price	End price	Volume traded	Total value of transaction
2011-1-7	13:37	8.22	8.22	8.21	8.22	16647	13671104
2011-1-7	13:38	8.23	8.23	8.22	8.23	1802	1482496
2011-1-7	13:39	8.23	8.25	8.23	8.25	4581	3774208
2011-1-7	13:40	8.25	8.25	8.23	8.23	4042	3329248
2011-1-7	13:41	8.24	8.26	8.24	8.25	4717	3891200
2011-1-7	13:42	8.25	8.26	8.24	8.25	2049	1690592
2011-1-7	13:43	8.25	8.25	8.24	8.24	1739	1434112
2011-1-7	13:44	8.25	8.25	8.22	8.22	7498	6174688
2011-1-7	13:45	8.22	8.22	8.2	8.2	4329	3551424
2011-1-7	13:46	8.2	8.2	8.19	8.19	4693	3846880

Table 5 Input Data Format

The price unit is CNY (Chinese Yuan).

There are four prices presented in this table: the start price at one minute, end price at one minute, lowest price and the highest price. The maximum differences between these four prices at any minute will not exceed 0.02 CNY. In addition, the error of prediction caused by the model's inaccuracy is much larger than that caused by inaccuracies of data input. Thus the "start price" is used as the input price in this system.

The figure of our input data is as follows:

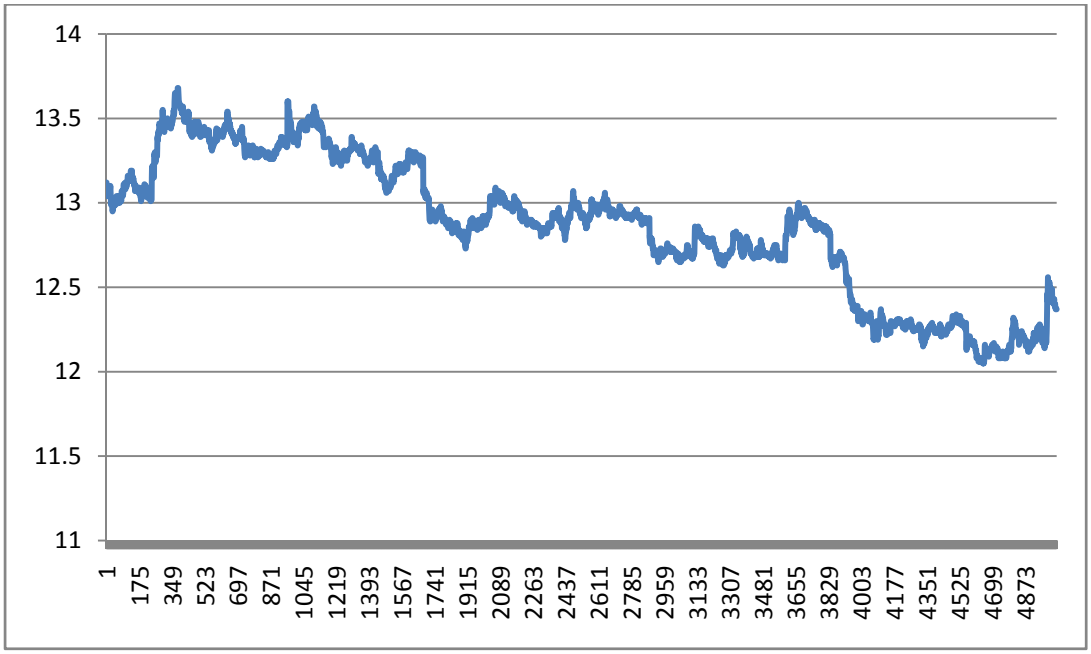

Figure 7 Data Example

The purpose is to select the optimal stock for trading. Considering the computational complexity and trading efficiency, eleven stocks have been chosen for analysis. The details of these stocks are as follows:

Name	Code	Sector
China National Petroleum Corporation	SH 601857	energy, petroleum
Shenhua Group Corporation	SH 601088	energy, petroleum
Industrial and Commercial Bank of China	SH 601398	finance
CITIC Securities	SH 600030	finance
PING AN Insurance Group	SH 601318	finance
Baosteel Group	SH 600019	steel
Aluminum Corporation of China	SH 601600	nonferrous metal
Shanghai Automotive Industry Corporation	SH 600104	automobile
Kwei Chow Moutai Corporation	SH 600519	wine
China Railway Group	SH 601390	railway
Air China	SH 601111	aviation
Dongfang Electric Corporation	SH 600875	electrical power

Table 6 Stock for Analysis

4.1.2 System Structure

The procedure employed to process the data is shown in figure 7.

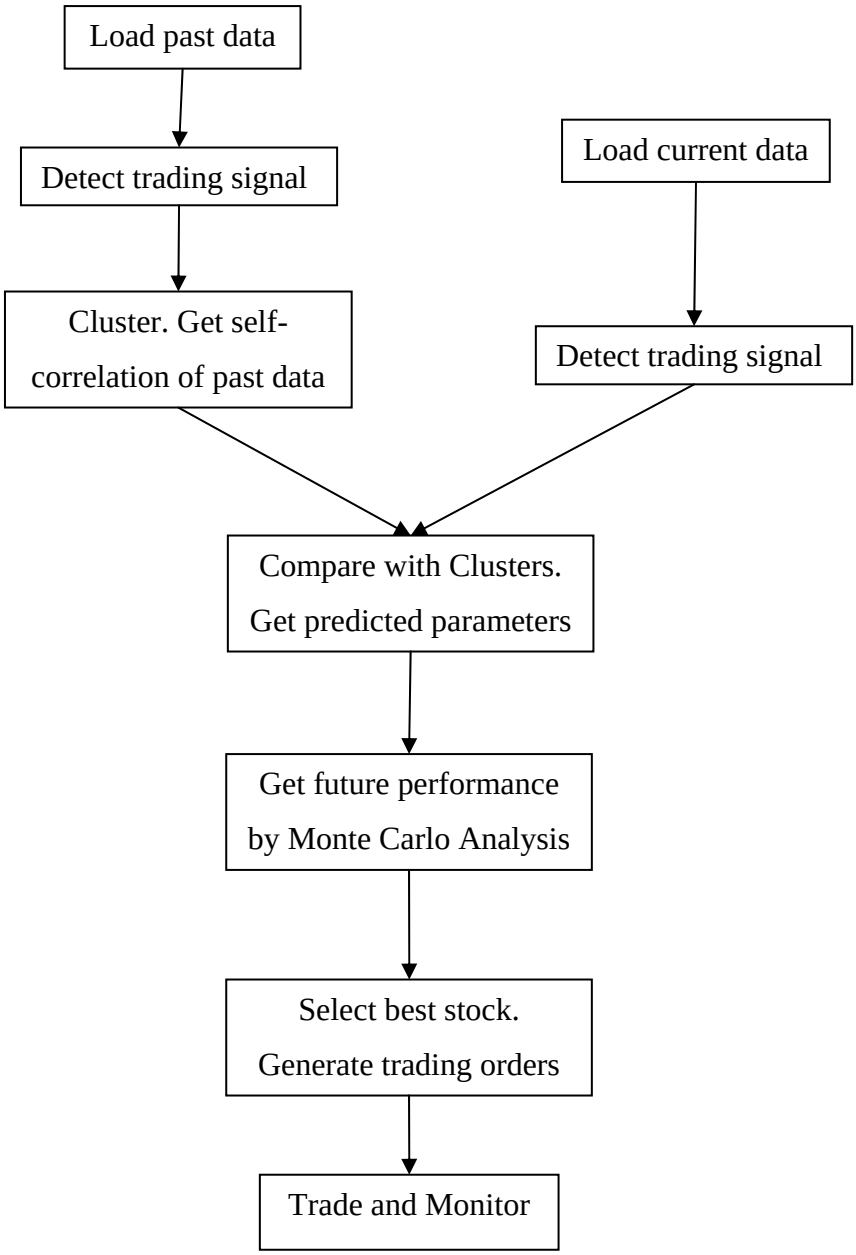

Figure 8 Data Processing Procedure

This data processing procedure can be separated into five major steps. The purpose of each step is shown in the following table.

Name	Function
Data_Process.cpp	1 Load stock data
Signal.cpp	1 Calculate parameters 2 Detect opportunity for trading in past stock data
Dbscan.cpp	1 Obtain clusters of each stock 2 Get self-correlation of 1000-minute long data before each trading signal
MC.cpp	1 Predict future expected rate of return after each trading signal occurs 2 Analyze each stock's future performance by considering both expected rate of return and volatility 3 Select the stock for trading whose future performance is the best 4 Output trading orders
Trade.cpp	1 Buy and short sell stocks according to trading orders from MC.cpp 2 Monitor the performance of each trade, determining when to end. 3 Calculate the trading result.

Table 7 Function of each Module

The detailed description of each function's implementation will be discussed in section 4.3.

4.2 Assumptions

4.2.1 Short term model

In the long term, highly nonlinear and irregularly-shaped data can be affected by many factors. In addition, no mathematical model can perfectly describe this system. However, in the short term, many studies [4] claim the log-normal model (equation 2.6) of stock prices is sufficient for analysis. In this thesis, the log-normal model has been chosen to characterize stock price data in short-term.

In addition, the model yields the future stock price that is dependent upon two parameters: the expected rate of return and volatility. The expected rate of return is caused by investors' response to new information, and the volatility is created by investors' speculation of this stock. Since investors' response to new information and their extent of speculation won't change minute by minute, we can assume the expected rate of return and volatility also won't change instantly. So in this thesis it can be assumed that these two parameters are constant throughout the five days.

4.2.2 Semi Efficient Market

There is disagreement between the Efficient Market hypothesis and the Behavioral Finance theory; however some empirical studies have been shown to support EMH. For this study the middle ground between these two theories is upheld meaning that the market is efficient throughout most events, while some patterns still exist. We have four reasons to support this assumption.

First, only a small fraction of investors realize the patterns are not likely to have a large affect on market activity. Since most successful investors prefer to keep their strategy a secret, the recognized patterns will not be likely to spread in the market. Thus the efficiency of the market can live with patterns realized by small number of investors.

Secondly, many investors cannot predict the market correctly due to limited math skills. Most popular models of stock analysis involve familiarity with stochastic mathematics, nonlinear regression and so on. Clearly, not all investors are aware of these complex mathematical principles.

Thirdly, the model of patterns may be time-varying. The model is based exclusively on the current market's performance, which further increases the difficulty of predicting future activity.

Thus the market won't be fully efficient because not all investors can understand the market 100% correct.

Lastly, empirical study that supports the Efficient Market Hypothesis only shows the positive aspects of this theory. No study can prove that EMH is one hundred percent correct. Since only a small departure from total accuracy can still lead to a huge profit, we cannot use the recent empirical study to deny the possible profiting chances from predicting the market.

4.2.3 Consistent Behavior

Under the assumption previously discussed in section 4.2.2, it is obvious that there are some sufficient methods to model the future movements of stock prices. From our

study of Behavioral Finance Theory, we know the fluctuations of stock prices are heavily based on how investors interpret the financial information presented to them. If events remain stable, then individuals' behavior is also stable. These "events" won't change frequently. Thus, we can assume that investors' behavior will be consistent over time.

There are some reasons to support this assumption: firstly, every investor forms a consistent way to process new information that confers consistent probability distributions about future rates of return; secondly, facing similar distribution, they make similar trading decisions, so they have a consistent decision-making process.

In this study, it can be assumed that investors have consistent trading behaviors that are reflected as time-series data with strong correlation. The purpose of this thesis is to find out what investors' behavior will be when similar trading signals comes up, thus the focus on uncovering the correlation of time series data.

4.3 Design and Algorithms

4.3.1 Data Pre Processing

The data pre processing function is implemented by "Data_Process.cpp.".Its main contribution is loading the past stock price data and converting them into an appropriate format.

Data from the 11 stocks are loaded into a vector, whose unit contains the information shown in Table 8.

Variable Name	Function
vector <Line>	time series data of stock price
string name	stock's name
int code	stock's code

Table 8 Content of vector <Stock>

The content in unit of <Line> is shown in Table 9.

Variable Name	Function
`int year`	year information
`int month`	month information
`int day`	day information
`int hour`	hour information
`int minute`	minute information
`float price`	stock price
`int volume`	trading volume at the minute
`int amount`	trading hands at the minute

Table 9 Content of vector <Line>

After being processed by "Data_Process.cpp," the time series data is ready for further analysis.

In this study, stock price data are used from 2011.1.4 to 2011.6.30 as experimental data and data from 2011.7.1 to 2011.7.29 are used as test data.

4.3.2 Detection of data peaks and valleys

Investors are particularly interested in “special events” of stock prices. When special events show up in stock price figures, investors usually start to think of making some changes to their investments. Since these “patterns” can trigger investors to make decisions, they are called “special data events” in this thesis. Clearly, there are many events. However, the most reliable and most predictable ones are the “peaks and valleys”. In this thesis, the predicting function is only triggered when a peak or valley appears and the self-correlation is only analyzed between data peaks and valleys.

A valley of data is the situation that the stock price keeps falling for some time then an increase comes up. It is shown in Figure 9 (from stock SH600875, July data).

Figure 9 Valley of Data

The left circle indicates the valley of a stock price. The second circlet shows an increasing after falling for a long time.

Similarly, a peak of data is shown in Figure 10.

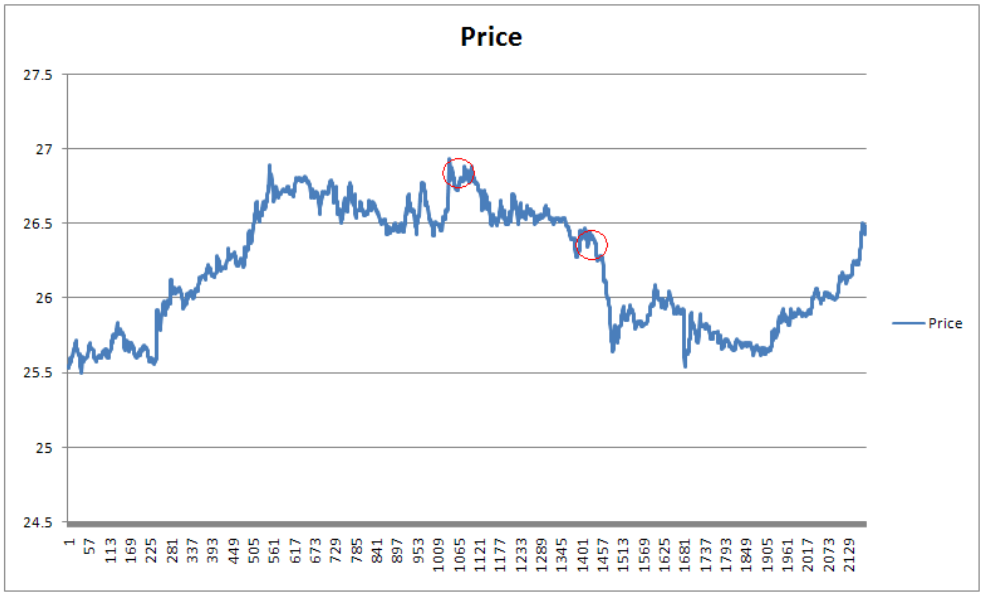

Figure 10 Peak of data

The left circle represents a stock price at its peak. The second circle indicates an falling after long term increasing

The peaks and valleys have been automatically detected by implementing "Signal.cpp," which outputs information whose content are shown in Table 10.

Variable Name	Function
`int month;  int day` `int hour;  int minute`	time of a peak or valley
`double price`	stock price of a peak or valley
`bool flag_max`	Indicate valley or peak. 1 peck, 0 valley
`bool Core_Flag`	Used in clustering. Indicate whether the point is core

Table 10 Contents

The algorithm used for detecting peaks and valleys is the sliding window. Take detecting valley of data as an example. The basic idea is to use a sliding window with appropriate length. If the maximum price in the window is in front of the minimum price and there is enough increase at the end of the window, then it can be concluded that the valley of data appears. The algorithm is shown in Figure 11.

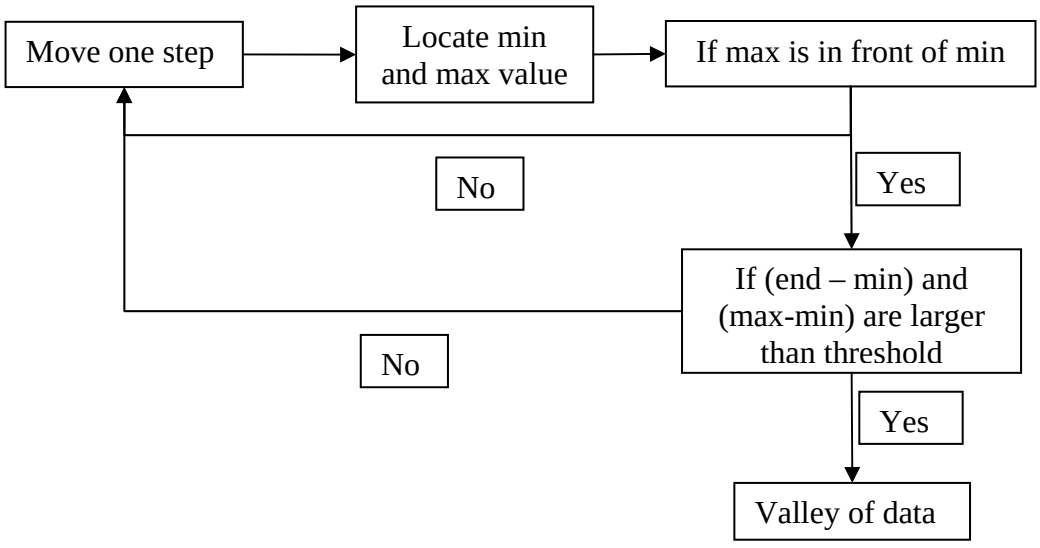

Figure 11 Valley of data Detecting Algorithm

There are four possible outcomes from the detecting of bottom signals. Three outcomes without a bottom signal are shown in Figure 12.

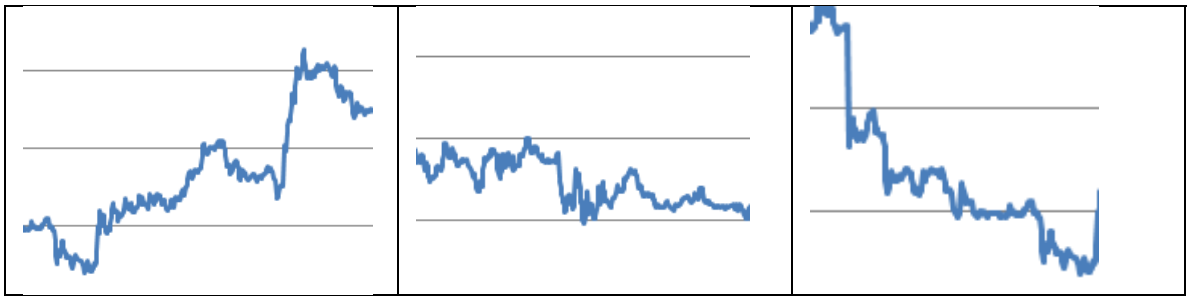

Figure 12 Signals without data valley

The three graphs from left to right show: 1 min value in front of max. 2 (max-min) doesn't exceed threshold. 3 not enough of an increase in stock price after a long-period of falling. A typical detection of bottom signals is shown in Figure 13.

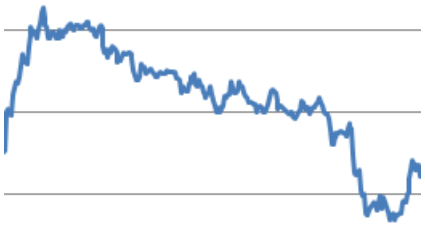

Figure 13 Detection of one valley

After implementing “Signal.cpp”, all peaks and valleys can be automatically located. An example is shown in Figure 14 taken from stock SH601318 from 2011.6.23 to 2011.7.8.

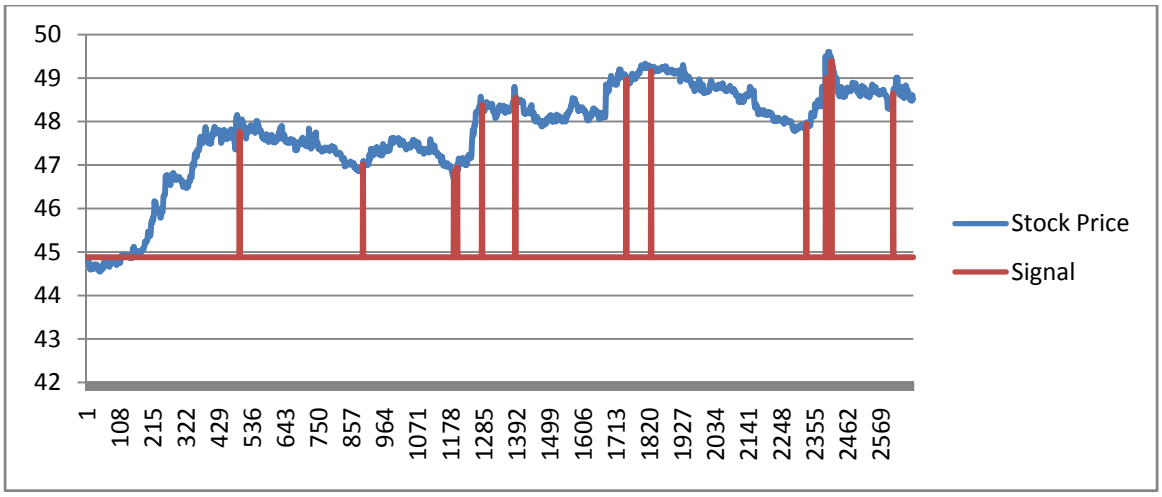

Figure 14 Peaks and Valleys Example

4.3.3 Clustering

Basically, the clustering method is used to obtain the self-correlation of time series data. If two data segments are similar enough, then they will be put in the same cluster. Thus, the data segments in the same cluster have a high correlation. This indicates that these two data segments are caused by similar investment behaviors. So the self-correlation of data segments can be analyzed by implementing the clustering method.

Stock price data is a coherent time series data. To analyze its self-correlation, the total time series must be broken down into relatively small segments. After that, it is feasible to compare and cluster these data segments.

This idea can be further illustrated by Figure 15. The data segments within the two red rectangles are the same length. We just cluster the data segments like these two in red rectangles.

Figure 15 Data Segment Example (SH601600, July)

It must be noted that all data segments used in this thesis end with a peak or valley. (The left data segment ends with a peak, while the right one ends with a valley.)

These data segments are vectors. Before comparing them, some modifications need to be done. Actually, of interest is at what segment the mode is located. For example, it

must be determined whether it is growing, decreasing or vibrating while its actual value is of no consequence. Some data segments have similar modes but very different values. This could be illustrated by Figure 16.

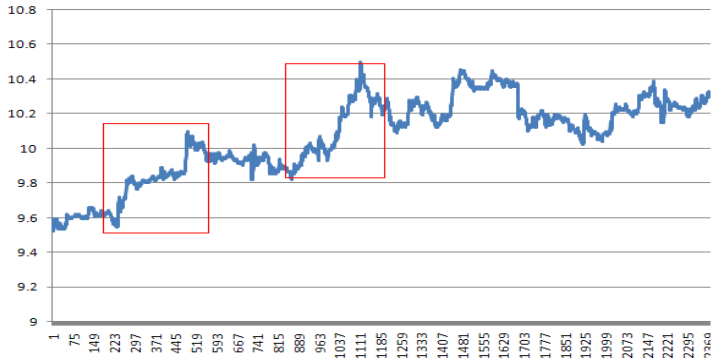

Figure 16 Similar Data Segments

The two data segments in the red rectangle have similar modes, since they are both growing. However, the Euclidean distance between them is large, because they have a static difference from the beginning.

If this static difference can be deduced, then data segments from similar mode can be very close in Euclidean distance without affecting their modes as shown in Figure 17.

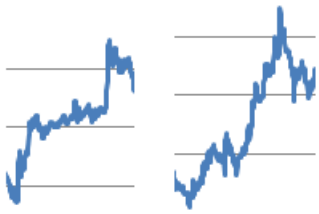

Figure 17 Data Segments after Modification

In this study, the difference between the starting values of two data segments is calculated. Then, this difference is deducted from the one with a larger starting value.

Also we use Euclidean distance to represent the distance between two data segments. The calculation is shown in equation (4.1).

Data Segment 1: $(x_1, x_2, ..., x_n)$

Data Segment 2: $(y_1, y_2, ..., y_n)$

Their distance is: $D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$ (4.1)

The clustering method used in this study is the density-based spatial clustering of applications with noise method (DBSCAN). Its basic purpose is to separate points in a high-density area from points in low-density area. Thus the cluster has a larger density than the outside.

DBSCAN classifies points into 3 categories:

1 Core Point: A core point has more than a certain number of points (MinPts) which are close enough (distance smaller than Eps) to it.

2 Board Point: A board point doesn’t have enough points which are close enough to it. However, it is close enough to a core point.

3 Noise Point: A point that is not a core point or a board point.

DBSCAN’s clustering policy is as follows:

1 If the distance of two core points is smaller than EPS, then they are placed in the same cluster.

A board point is placed in the same cluster with its core point

3 Discard all noise points.

The code is shown in Appendix B.

The reason we use the DBSCAN clustering method is because we have little knowledge about the clusters of our data. We do not know how many clusters are there in advance. However, DBSCAN doesn’t require this information. Thus a good clustering accuracy can be obtained without knowing the number of clusters.

In this last section, we illustrate some data segments which are placed in the same cluster by the DBSCAN method, as seen in Figure 18.

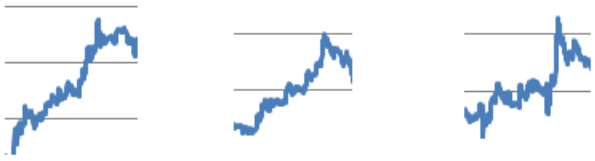

Figure 18 Data segments from same cluster

4.3.4 The Monte Carlo Simulation

The expected rate of return can be obtained by analyzing the self-correlation of data segments, but not the future performance of stock prices. In addition, a stock's volatility plays an important role in the stock's future performance. This can be illustrated by Figures 19 and 20.

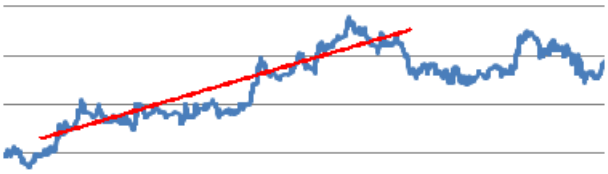

Figure 19 High expected rate of return with low volatility

There can also be a data segment with a low expected rate of return but a high volatility, as depicted in Figure 20.

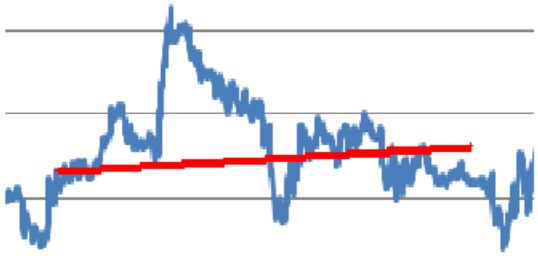

Figure 20 Low expected rate of return with high volatility

A stock with a high predicted expected rate of return is very likely to yield a higher price in the long run. However, a stock with a low expected rate of return but large volatility may temporarily fluctuate into a higher price several times in a short amount of time. Technically, there is a trade-off between the volatility and expected rate of return. The stock's future performance is predicted by considering both volatility and expected rate of return comprehensively.

In this study, we predict the stock's future performance by the Monte Carlo Simulation. The steps of Monte Carlo method are as follows:

1 Predict the parameters: future expected rate of return, volatility and correlation of stocks.

2 Simulate the stock's future performance by employing equation (4.2) for N times. Calculate N sequences of future stock price. Then sort these prices.

$$InS_T \sim \phi(InS_0 + (\mu - \frac{\sigma^2}{2})T, \sigma^2 T) \qquad (4.2)$$

3 Take the average M highest prices of the sorted N sequences as the price representing the stock's future performance

Predicting Expected Rate of Return:

After a new data segment ending with a peak or valley has shown up. Then this data segment is compared with the existing clusters. The Euclidean distances for each data segment in each cluster are calculated. The minimum distance is the "distance" between the current data segment and the cluster.

All the distances between the current data segments and the clusters are calculated. The cluster that is closest to the current data segment is called the current data segment's "closest cluster."

Therefore, the growth rates after the trading signal, of each data segment in the "closest cluster" are calculated, as shown in Figure 21.

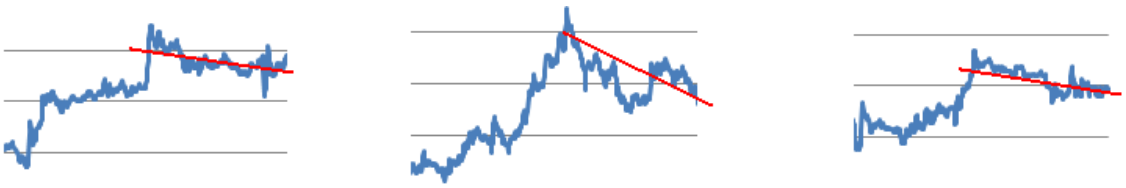

Figure 21 Data segments in one cluster

Figure 212 illustrates the current data segment ending with a peak. Its closest cluster is shown in Figure 21.

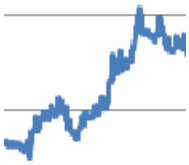

Figure 22 Current data segment

Figure 21 shows the current data segment ends with a peak signal. Its closest cluster is shown in Figure 20.

Then the predicted expected rate of return of the current data segment is the average slope of the three red lines, shown in Figure 21.

We calculate the volatility by the following method. Define:

$n+1$: Number of observations.

S_i: Stock price at end of ith interval, with $i = 0,1,.....,n$

σ: Volatility of stock price per minute

and let

$$u_i = In(\frac{S_i}{S_{i-1}}) \text{ for } i = 1,2,.....,n \qquad (4.3)$$

Then the volatility of the stock price each minute is estimated as:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} u_i^2 - \frac{1}{n(n-1)}(\sum_{i=1}^{n} u_i)^2} \qquad (4.4)$$

To calculate the correlation between two stock price data, we define:

σ_x: Volatility of stock x

σ_y: Volatility of stock y

X , Y : Stock prices

$n+1$: Number of observations.

ρ: Correlation between two stocks

cov: Covariance of daily changes in stock x and y

and let:

$$x_i = \frac{X_i - X_{i-1}}{X_{i-1}}, \qquad y_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}} \tag{4.5}$$

Then the correlation is calculated as:

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} x_i y_i}{n \sigma_x \sigma_y} \tag{4.6}$$

The Monte Carlo simulation requires two correlated standard normal distributions that are generated by the following procedure.

First, generate two independent variables U and V from uniform distribution on (0,1).

$$x = \sqrt{-2InU}\cos(2\pi V) \tag{4.7}$$

$$y = \sqrt{-2InU}\sin(2\pi V) \tag{4.8}$$

Then x and y are two independent standard normal distributions. Let:

$$\phi_1 = x \tag{4.9}$$

$$\phi_2 = \rho x + y\sqrt{1-\rho^2} \tag{4.10}$$

ϕ_1 and ϕ_2 will be two standard normal distributions with correlation ρ.

Lastly, the best stock is selected for trading whose future simulated performance is the highest.

4.3.5 Trading and Monitoring

After performing the above calculations, the performance of each trade needs to be monitored in order to maximize the profit and control the risk. Two boundaries should be set, the stop loss boundary and the maximum profit boundary. In addition, a maximum available time for each trade should also be set. If the time is exceeded, the trade will be terminated and the stocks will be sold. This phenomenon is explained in the following examples using the buy orders. We can have three scenarios that are shown in the following figures.

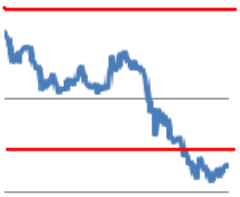

Figure 23 Stock price hits stop loss boundary

The stock price may fall after purchase. To prevent greater loss in the future, a stop loss boundary has been set after which point the stock will be sold

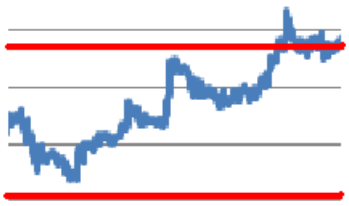

Figure 24 Stock price hit maximum profit boundary

After the stock price increases to a high enough value, it may continue to increase or fall in value. To take advantage of the best time to sell, the maximum profit boundary has been set. If the stock price hits that boundary, the stock will be sold.

Figure 25 Stock price escapes both boundaries

Moreover, the position cannot be held forever. If the time is up, the position should be sold whether it reaches its boundary or not.

Chapter 5: Results

5.1 Trading Results

The July 2011 data was used as the test data. The trading results are shown in Table 11.

Date	Result
7.1	7.4 10:9 BUY start:18.56 end:19.5 profit:178600
	7.7 13:59 SHORT start:30.93 end:31.09 profit:-1600
7.4	7.8 14:51 BUY start:192.47 end:194.2 profit:328699
	7.8 14:21 SHORT start:9.83 end:10.15 profit:-3200
7.5	7.11 11:4 BUY start:190.3 end:195.3 profit:950000
	7.11 10:15 SHORT start:6.17 end:6.02 profit:1500
7.6	7.12 10:28 BUY start:19.01 end:18.23 profit:-148200
	7.12 9:31 SHORT start:49 end:47.48 profit:15200
7.7	7.14 9:49 BUY start:18.66 end:18.18 profit:-91199.9
	7.8 10:7 SHORT start:10.43 end:10.11 profit:3200.01
7.8	7.14 11:29 BUY start:193.45 end:193.44 profit:-1898.96
	7.14 13:22 SHORT start:4.34 end:4.36 profit: 200
7.11	7.15 11:0 BUY start:25.77 end:26.15 profit:72199.8
	7.12 10:8 SHORT start:10.42 end:10.1 profit:3200
7.12	7.18 13:30 BUY start:30.14 end:30.13 profit:-1900.04

Continued

Table 11 Trade Result

Table 11 continued

7.13	7.19 10:50 BUY start:193.9 end:199.78 profit:1117200
	7.19 10:18 SHORT start:13.05 end:12.65 profit:4000.01
7.14	7.21 9:58 BUY start:193.5 end:203.6 profit:1919000
	7.20 10:23 SHORT start:47.58 end:47.41 profit:1700.02
7.15	7.21 10:48 BUY start:30 end:29.48 profit:-98800.1
	7.21 10:13 SHORT start:47.93 end:46.45 profit:14800
7.18	7.22 11:20 BUY start:18.18 end:18.01 profit:-32300
	7.21 9:41 SHORT start:48 end:46.5 profit:15000
7.19	7.22 11:26 BUY start:197.85 end:207.75 profit:1881000
	7.25 10:48 SHORT start:10.76 end:10.51 profit:2500
7.20	7.26 13:59 BUY start:25.67 end:24.32 profit:-256500
	7.26 13:26 SHORT start:200.1 end:202.75 profit:-26499.9
7.21	7.27 14:37 BUY start:25.35 end:24.5 profit:-161500
	7.27 14:59 SHORT start:204 end:204.5 profit: 5000
7.22	7.28 10:12 BUY start:10.63 end:10.78 profit:28499.9
	7.28 14:4 SHORT start:207.51 end:204.35 profit:31599.9
7.25	7.29 15:0 BUY start:203.4 end:207.18 profit:718200
	7.25 14:45 SHORT start:10.05 end:9.72 profit:3300
7.26	7.29 15:0 BUY start:202.85 end:207.18 profit:822697
	7.27 9:37 SHORT start:10.42 end:10.06 profit:3600
7.27	7.29 15:0 BUY start:24.31 end:24.47 profit:30400
	7.29 15:0 SHORT start:10.73 end:10.65 profit:799.999
7.28	7.29 15:0 BUY start:202.69 end:207.18 profit:853098
	7.29 15:0 SHORT start:10.73 end:10.65 profit:799.999
7.29	7.29 15:0 BUY start:24.38 end:24.47 profit:17100
	7.29 15:0 SHORT start:44.88 end:44.5 profit:3800.01

Take the first line of 7.1's result as an example. The meaning of "7.4 10:9 BUY start:18.56 end:19.5 profit:178600" is as follows:

1 Buy order executes in 7.2

2 The sell time of the trade is 7.4 10:09AM

3 The buying price is 18.56 CNY.

4 The selling price is 19.5 CNY.

5 The total profit for this transaction is 178600 CNY.

5.2 Statistical Results

The average return rate of each trade is 0.0206021, and the return rate in July was 0.0715761.

The comparison between the return rate which is derived from the method proposed by this study and the major Chinese mutual fund are shown in Table 12.

Trade period 7.1.2011-7.29.2011			
Fund	Start Value	End Value	Return Rate
Bank of Communications Schroders Fund	1.0260	0.99	-0.03509
China Southern Fund	1.0090	1.0101	0.00109
China Asset Management Fund	3.891	3.906	0.00386
Orient Fund	1.0000	0.9910	-0.00900
ICBC Fund	1.3709	1.4101	0.02859
China Merchants Fund	1.0130	0.9980	-0.01481
SYWG BNP Paribas Asset Management Fund	1.0390	1.0330	-0.00577

Continued

Table 12 Fund Return Rate

Table 12 continued

Penghua Fund	1.0000	0.9860	-0.01400
E Fund	1.3400	1.3280	-0.00896
China Universal Asset Management	0.8920	0.9280	0.04036
Harvest Fund	1.0830	1.0710	-0.01108
Bosera Fund	0.7740	0.7610	-0.01680
Bank of China Fund	1.0590	1.0600	0.00094

We can see that our method performs better than most Chinese stocks.

The predictive accuracy is 31/41, which is equal to 75.6%. The predictive accuracy of the major models is shown in Table 13.

Prediction Model	Trend Prediction Accuracy
Time series motif	83%
BP Neural Networks	72%
Association Rules	75%
Dbscan	75.6%

Table 13 Predictive Accuracy

At first glance, our method doesn't seem more accurate than the other predictive methods. However, our method only predicts the trend at the turning point, which increases the difficulty of prediction. Since the other predictive models tend to keep their predictive trends consistent with the current trend, our model is actually much more

accurate at the turning point. More importantly, our model is much more practical for the financial industry.

Chapter 6: Conclusion and Suggestions for Future Work

6.1 Conclusion

This thesis proposes a method for analyzing the self-correlation of nonlinear, irregularly-shaped time series data. We used the stock price data as the example and employed DBSCAN clustering as the method of analyzing the self-correlation.

The future parameter of the data was predicted based on the self-correlation between the current and past data segments. After that, the Monte Carlo method was used to simulate the future performance of each stock based on previously predicted parameters. The stock with the best future performance was selected for trading. Lastly, the trading orders were implemented and monitored

The method proposed in this thesis has four advantages:

1 It highly reduces the computational cost of prediction. It may make the algorithm trading system possible on laptops.

2 Synergy is created between DBSCAN and the Monte Carlo method, which creates a more efficient stock selecting strategy.

3 The predictive accuracy at the “turning point” of times series data is highly increased.

4 It allows for a higher investment return, which will be highly sought after in the financial industry.

6.2 Future Work

There are many types of nonlinear, irregularly-shaped time series data which represent similar patterns repeated over time. The method proposed in this study can also be implemented in predicting the future trend of the data. The most famous examples are as follows:

1 The price of other financial assets may be predicted, such as commodity futures, ETF and so on.

2 The usage of electric power in a particular area may also be predicted, which could be part of a “Smart Grid” project to increase the efficiency of the grid.

References

[1] Cox, D.R., and H.D. Miller. *The Theory of Stochastic Process*. London: Chapman & Hall, 1977.

[2] Cootner, P.H. (ed.) *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press, 1964.

[3] Neftci, S. *Introduction to Mathematics of Financial Derivatives*, 2nd edn. New York: Academic Press, 2000.

[4] Fama, E.F., "The Behavior of Stock Market Prices," *Journal of Business*, 38 (January 1965): 34-105.

[5] Kon, S. J., "Models of Stock Returns —— A Comparison," *Journal of Finance*, 39 (March 1984): 147-65

[6] Lye, Jenny, and Vance L. Martin, 1994, Towards a Theory of Nonlinear Models, in *Chaos and Nonlinear Models in Economics: Theory and Applications,* ed. John Creedy and Vance L. Martin; Elgar, Brookfield, 70-86.

[7] Engle, Robert F., 1982, Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation, *Econometrica*, v50, 987-1007.

[8] De Long, James Bradford, and Lawrence H. Summers, 1986, Are Business Cycles Symmetrical?, *The American Business Cycle: Continuity and Change*, NBER Studies in Business Cycle Series, v25, University of Chicago Press, 166-78.

[9] John Y. Campbell and Robert Shiller, "Stock Prices, Earnings and Expected Dividends," *Journal of Finance* 43 (July 1988), pp. 661-76.

[10] Nicholas Barberis and Richard Thaler. *The Handbook of the Economics of Finance*, Amsterdam: Elsevier, 2003

[11] Ruben D. Cohen (2002) "The Relationship between the Equity Risk Premium, Duration and Dividend Yield," *Wilmott Magazine*, pp 84-97, November issue.

[12] Maurice Kendall, "The Analysis of Economic Time Series, Part Ⅰ: Price," *Journal of the Royal Statistical Society* 96 (1953)

[13] W.F.M. De Bondt and R. H. Thaler, "Do Security Analysis Overreact?" *American Economic Review* 80 (1990). pp. 52-57

[14] Meir Statman, "Behavioral Finance," *Contemporary Finance Digest* 1 (Winter 1997), pp. 5-22

[15] Melanie Bowman and Thom Hartle, “Dow Theory,” Technical Analysis of Stocks & Commodities Vol. 8, No.9 (Sept. 1990)

[16] John A. Hartigan, *Clustering Algorithms*, 99th. New York: John Wiley & Sons, Inc., 1975, ISBN:047135645X

[17] Zvi Bodie, Alex Kane and Alan J. Marcus, Investments, 8th edition, Singapore: McGraw Hill, 2009, ISBN: 978-007-126325-2

[18] Shanghai Stock Exchange (1990-2011). “Trading Rules of Shanghai Stock Exchange”. http://www.sse.com.cn/sseportal/en/c01/c09/c01/p1075/c15010901_p1075.shtml

[19] Rachlin, G, Last, M, Alberg, D, Kandel, A 2007, 'ADMIRAL: A data mining based financial trading system', in Proceedings of the IEEE Symposium on Computional Intelligence and Data Mining, Eds. Piuri, V, Fogel, D, Bonissone, P & Yen, G, Hawaii, U.S.A, pp. 720-725.

[20] Huacheng Wang, Yanxia Jiang, Hui Wnag, 2009, Stock Return Prediction Based on Bagging-Decision Tree, Proceedings of The 2009 IEEE International Conference on Grey Systems and Intelligent Services, Nov, pp:10-12.

[21] Jiang, Y.F. , C.P. Li and J. Z. Han, Stock Temporal Prediction Based on Time Series Motifs, International Conference on Machine Learning and Cybernetics, 2009.

[22] Kuang Yu Huang, J. Chuen-Jiuan, Ting-Cheng Chang: A RS model for stock market forecasting and portfolio selection allied with weight clustering and Grey System theories. IEEE Congress on Evolutionary Computation 2008: 1240-1246.

[23] Jianfei Wu, Anne Denton, Omar el Ariss, Dianxiang Xu. Mining for Core Patterns in Stock Market Data. In Proceedings of ICDM Workshops'2009. pp.558~563.

[24] Preeti Paranjape-Voditel, Umesh Deshpande, "An Association Rule Mining Based Stock Market Recommender System," eait, pp.21-24, 2011 Second International Conference on Emerging Applications of Information Technology, 2011.

[25] Li Ping, Xing Wenjing, Huang Guangdong, "Financial Asset Price Forecasting Based on Intertransaction Association Rules Mining," icee, pp.1422-1425, 2010 International Conference on E-Business and E-Government, 2010.

[26] Romain Pierrot, Hongyan Liu, "The Influence of Volume and Volatility on Predicting Shanghai Stock Exchange Trends," fskd, vol. 1, pp.470-474, 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008.

[27] I. Matsuda. *Application of Neural Sequential Association to Long-Term Stock Price Prediction*. Proceedings of IJCNN'91, Singapore, 1991. 7.

[28] J.H.wang and J.Y.Leu, “stock market trend prediction using ARIMA-based neural network,”Proc. Of IEEE conference on neural networks, vol.4, pp.2160-2165, 1996.

[29] Lean Yu, Huanhuan Chen, Shouyang Wang and K. K. Lai. Evolving Least Squares Support Vector Machines for Stock Market Trend Mining. IEEE Transactions on Evolutionary Computation. Vol 13, No. 1, Feb 2009.

[30] M. N. Nguyen, U. Omkar, D. Shi, J. B. Hayfron-Acquah. Stock Market Price Prediction using Cyclic Self-Organizing Hierarchical CMAC. In Proceedings of ICARCV'2006. pp.1~6.

[31] Zhou Yixin, Jie Zhang, "Stock Data Analysis Based on BP Neural Network," iccsn, pp.396-399, 2010 Second International Conference on Communication Software and Networks, 2010.

[32] Wuthrich, B., Cho, V., Leung, S., Sankaran, K., and Zhang, J.: *Daily Stock Market Forecast from Textual Web Data*. 1998 IEEE International Conference on Systems, Man and Cybernetics.

[33] Fung, G.P.C.; Yu, J.X.; Lam, W.: Stock Prediction: Integrating Text Mining Approach Using Real-time News. In: Proceedings IEEE Int. Conference on Computational Intelligence for Financial Engineering. Hong Kong 2003, pp. 395-402.

Appendix A: ITO's Lemma and Derivation of the Lognormal Property

A variable x follows the Ito process, if the following condition is met:

$$dx = a(x,t)dt + b(x,t)dz \tag{A.1}$$

$a(x,t)$ and $b(x,t)$ are functions of x and t. dz is z Wiener process. x has a drift rate of a and a variance rate of b^2. Ito's Lemma states that the function $G(x,t)$ follows the following process:

$$dG = (\frac{\partial G}{\partial x}a + \frac{\partial G}{\partial t} + \frac{\partial^2 G}{2\partial x^2}b^2)dt + \frac{\partial G}{\partial x}bdz \tag{A.2}$$

dz is the same Wiener process as in (A.1).

Then $G(x,t)$ also follows an Ito process, with a drift rate of:

$$\frac{\partial G}{\partial x}a + \frac{\partial G}{\partial t} + \frac{\partial^2 G}{2\partial x^2}b^2 \tag{A.3}$$

and a variance rate of

$$(\frac{\partial G}{\partial x})^2 b^2 \tag{A.4}$$

Now we derive the lognormal property of stock price by Ito's Lemma.

The stock price follows:

$$dS = \mu S dt + \sigma S dz \tag{A.5}$$

We define:

$$G = InS \tag{A.6}$$

From equation A.2 we know:

$$dG = (\mu - \frac{\sigma^2}{2})dt + \sigma dz \tag{A.7}$$

This indicates that the increment of G is normally distributed. That is:

$$InS_T - InS_0 \sim \phi[(\mu - \frac{\sigma^2}{2})T, \sigma^2 T] \tag{A.8}$$

This is just the lognormal property of stock price.

Appendix B: Code of DBSCAN

```
Dbscan::Get_Cluster( int stock_code, bool flag)

{ for( i=0 ; i<size ; i++ )

  { if( status[i] == unclassified )

  {if( Expand_Cluster(stock_id, i, Cluster_Id, distance, status, work_data, flag, Eps[i] ) )

       {Cluster_Id ++ ;}

    }

}}



Dbscan::Expand_Cluster(int stock_id, int Point_Id, int Cluster_Id,

vector<vector<double>> distance, vector<int>& status, vector<Extremum>& work_data,

bool flag, double Eps)

while( ! Neighbour_Point.empty() )//while the neighbour set is not empty

{//current point is the first point of Point's remaining neighbours

  Current_Point = Neighbour_Point[0];
```

```
      Neighbour_Point.erase( Neighbour_Point.begin() );
      Neighbour_Current_Point.clear();
      //get Current_Point's neighbours
      for( i=0; i<size ; i++ )
      { if( distance[Current_Point][i] <=  Eps )
        {Neighbour_Current_Point.push_back(i);
          }}
    //if Current_Point is core
if( Neighbour_Current_Point.size() >= MinPts )
{ work_data[ Current_Point ].Core_Flag = TRUE;//mark Current_Point as 'Core'
    for( i=0 ; i<Neighbour_Current_Point.size() ; i++ )
    { if( status[Neighbour_Current_Point[i]]==unclassified ||
status[Neighbour_Current_Point[i]]==noise )
    { if( status[Neighbour_Current_Point[i]] == unclassified )
    { Neighbour_Point.push_back( Neighbour_Current_Point[i] );
      }
```

```
    Cluster_Solo.push_back( work_data[ Neighbour_Current_Point[i] ] );

      status[Neighbour_Current_Point[i]] = Cluster_Id;

    }

    }

}

}
```