Using Pattern Recognition for Investment Decision Support in Taiwan Stock Market

An-Pin Chen, Yi-Chang Chen, Chi-Pin Cheng, Ju-Yin Lin
Institute of Information Management
National Chiao Tung University
Hsinchu, Taiwan
886-3-5712121 ext 57425
{apc, ycchen, abinbin, pika}@iim.nctu.edu.tw

Abstract

In Taiwan stock market, it has been accumulated large amounts of time series stock data and successful investment strategies. The stock price, which is impacted by various factors, is the result of buyer-seller investment strategies. Since the stock price reflects numerous factors, its pattern can be described as the strategies of investors.

In this paper, pattern recognition concept is adapted to match the current stock price trend with the repeatedly appearing past price data. Accordingly, a new method is introduced in this research that extracting features quickly from stock time series chart to find out the most critical feature points. The matching can be processed via the corresponding information of the feature points. In other words, the goal is to seek for the historical repeatedly appearing patterns, namely the similar trend, offering the investors to make investment strategies.

Keywords: Pattern Recognition, Taiwan Stock Market, Time Series, Feature Extraction

1. Introduction

Stock exchange is influenced by many factors, including politics, economics, international statuses, significant news, and so forth. The composite result will be displayed on the market, and thus the stock price volatility is reflected. Since the stock price is the final consequence of various components, it is more meaningful to present the variation of stock price with the chart than with numerical data. Hence, stock trend chart is one of the most useful tools for stock analysis [2][7][9][10].

In recent years, Behavioral Finance has been getting more and more attention. When one is making strategies, he, who is affected by mental elements, is not always rational. Certainly, daily stock price corresponds to the current circumstances. Thereafter, stock data mining [6], which obtains knowledge through relevant analysis, reveals that investors make investment strategies according to each factor.

Consequently, using the objective approach in this paper discovers investment behavior patterns — the curves of stock price trend. First, applying feature extraction to stock time series to extract the most critical feature points. Then, do the matching of the relevant information among the feature points. The proposed new pattern recognition method can substantially improve the efficiency of pattern matching about time series data. Thus, it is useful for investors to make decisions.

2. Literature Review

2.1 Stock Market Volatility Factors

Stock price is affected by some factors [5], including stock market, industry, and corporation factors. Stock market factors involve macroeconomics, international statuses, and domestic politics. Industry factors consist of industry conditions, business life cycle, and law measures. Corporation factors concern with dividends, operating performance, firm structure and restructuring. To sum up, stock price is influenced by all factors; either corresponding to long-term or short-term trend.

Based on the self-similarity of fractal theory [9], the shape will not be changed even if a segment of long-term trend is either magnified or minified. In another word, the shape of long-term stock price trend is similar with short-term. The fluctuation of short-term trend will not be taken into consideration while observing long-term trend. In the same way, the volatility of ultra-short term trend will be neglected while focusing on short-term trend. For example, the minute stock price volatility will not be taken into account when considering daily stock price data. Therefore, it is necessary to find out quickly the critical feature points to represent the original curves.

Since the volatility factors are subjectively defined and explained by investors, investment will correspond to individual character and mental factors; that is, there is no objective standard. Thereupon, in this paper, an objective approach is applied to pattern discovery of repeated behaviors by the historical curves.

2.2 Time Series Feature Extraction

The reason for applying feature extraction to stock time series is because the preceding matching approaches are time-consuming. The most typical models are Hidden Markov Model (HMM) [1] and Dynamic Time-Warping (DTW) [4]. The parameters of HMM require being trained; the training process takes a lot of time. Though the parameters of DTW do not need to be trained, the matching process is rather time-consuming.

The proposed Perpetually Important Point (PIP) approach [3], which is suitable for time series feature extraction, indicates that major or minor volatility may be included in the stock time series. Discovering the critical

feature points ensures to filter the insignificant minor volatility. Thus, the complete time sequence can be expressed in terms of a small number of feature points. Let P denote a time series, where the length is N. The purpose is to find out m PIPs, which works as follows:

Step1: The adjacent two PIPs are defined on the two ends of time series; in other words, the start point p_1 and end point p_2.

Step2: The third PIP p_3 is the point in P with maximum distance to the first two PIPs. The maximum distance denotes the vertical distance between a point p_3 and a line connecting the two adjacent PIPs (See Fig. 1), that is, $d = y_3 - y_c$.

Step3: The process continues searching for the next PIP in P with maximum distance to its two adjacent PIPs until sufficient m PIPs are detected, and m is predefined.

The pseudo code of the PIP approach process is described in Figure 2.

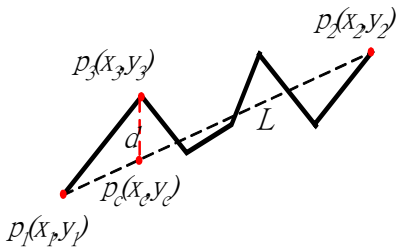

Figure 1. Time series feature extraction.

The Perceptually Important Point (PIP) approach is adapted to find out important features for human visual identification, and two approaches in extracting PIPs are provided. The first one is to find out the most significant feature points on the complete time series. The characteristics of second one is that the time interval is approximatively equal after extracting PIPs. See Figure 3 (1st PIP approach), and Figure 4 (2nd PIP approach), respectively.

The second approach is developed to find out two adjacent PIPs with maximum time interval. After the step, it is found that the each interval between two adjacent PIPs is the approximative distance. The two feature extraction approaches will lead to different results, even with the same data.

```
Function PIP(P,m)
  Input: sequence p[1…N]
  Output: pattern pip[1…m]
Begin
  Set pip[1]=p[1], pip[N] = p[m]
  Repeat until pip[1…m] all filled
    Select point p[j] with maximum distance to the adjacent point
    in pip
    Add p[j] to pip
End
```

Figure 2. Pseudo code of PIP[3].

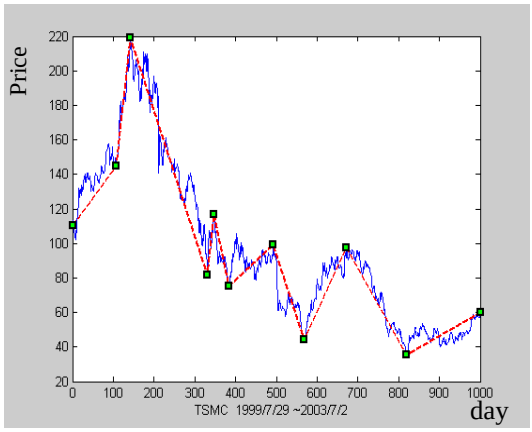

Figure 3. PIP Method 1. The feature points are represented by square points, the daily closing price are represented by solid lines, and the feature points are connected by dotted lines.

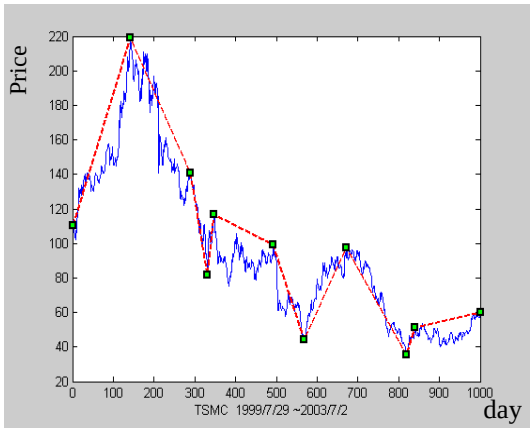

Figure 4. PIP Method 2. The feature points are represented by square points, the daily closing price are represented by solid lines, and the feature points are connected by dotted lines.

The two methods are chosen according to the problem domain. Here, the first method is preferred being applied to stock market data. The number of needed feature points can be given in advance. Since matching the original data is time-consuming; therefore, detecting the meaningful points by filtering detailed information can improve the matching work.

Let n denote the points on time series. The whole data has to be scanned once as long as seeking for one feature point. Hence, n^2 times of calculations are computed when detecting n PIPs. In real life, one usually keeps his eye only on several critical points, such as starting point, termination point, peak, valley, turning point, etc.

Therefore, only n $(n<<N)$ critical features are required. For this reason, n times of PIP approach calculations are computed when detecting time series, such that the time complexity will be $O(n \times N)=O(N)$. The other non-feature points can be neglected for being viewed as slight volatility.

3. Architecture

In this research, the data resource is derived from Securities & Futures Institute, where daily closing price is used to do the feature extraction by means of PIP approach firstly. Then, the angle and relative length are computed for matching similar patterns, namely resembling stock price trend (See Fig.5). Finally, the investment suggestion is given by the analysis.

3.1 Data Collection

Taiwan stock market is taken for example in this paper. The data originates from Securities & Futures Institute, where daily closing price is collected from 1994/1/5 to 2003/7/16.

3.2 Time Series Feature Extraction

The PIP approach is used to apply feature extraction to stock data because one will notice only a few significant feature points within a time period.

3.3 The Relation of Features

Calculate the slope of two adjacent feature points. Slope matching is complicated because slope lies within the range $[-\infty, \infty]$. Therefore, conversing the slope to the angle, and then using the angle to do the matching. There are some advantages of angle matching:

1. The different stock price (Y axis) can be matched with varied stocks. Hence, the stock price data needs to be normalized. For example, stock *A* falls within 10 to 30 dollars whose spread is 20 dollars. Stock *B* falls within 50 to 80 dollars whose spread is 30 dollars. The scale of stock *B* requires being reduced two-thirds, while stock *A* being magnified one and one-second can the angle matching be done.
2. Time axis (X axis) does not require being normalized. This is because there is no impact on two angles regardless of the length of time. As long as the number of the chosen feature points is equal, the data of M-day can be matched with that of N-day. For instance, 50 days of data can be matched with that of 60 days.

An error range ε is determined when matching, and the angles lie within the range $[-90^o, 90^o]$. Let α denotes the angle of a feature point such that $\alpha \pm \varepsilon$ are regarded as the identical angles (See Fig. 6). The smaller of the error range, the more precise of the matching result; otherwise, the more similar patterns can be matched.

The time interval of two feature points may differ in various patterns, which contribute to divergent implications applied to stock market data. Therefore, the distance between feature points should be considered when matching. In comparison of different lengths of time series flexibly, let the relative time length be the measure through dividing the time interval of two adjacent feature points by the total length of time series (See Fig. 7). In Figure 7, the relative length between point 1 and point 2 is *a/N*. The error of relative length is set not to be more than β when matching stock price time series.

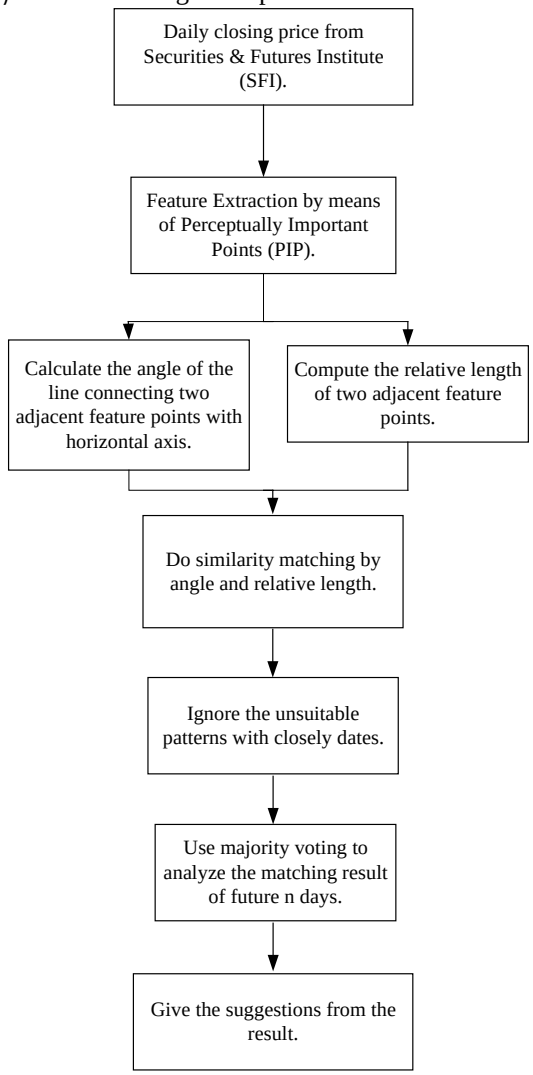

Figure 5.Research structure.

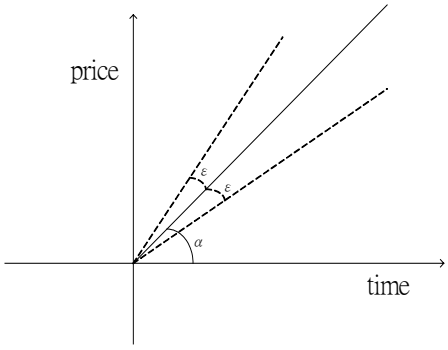

Figure 6. The matching angle is α, and the range of tolerance is ε.

3.4 Similarity Matching

The process of matching begins with finding out the most critical features by way of feature extraction; furthermore, compute the angle of feature point connecting to horizontal axis and relative length, described in section 3.3, for matching. First of all, angle

matching is employed with adjacent feature points. After matching the identical angles, the relative length is further matched. The matching patterns may be dissimilar even with the identical angle; that is why the relative length proceeds to be matched. The error of angle and relative length require being within the range of tolerance such that the matching of time series is regarded as identical. The stricter the making rules of similarity, the better resulting matching; nevertheless, the fewer times of matching the identical patterns. On the contrary, the looser the threshold, the matching result will be worse; yet, the more times of matching the identical patterns.

It is found in the experiment that the stocks of the same industry within the similar time period will usually result in identical trend. Therefore, finding out the same trend with different time period will be more meaningful.

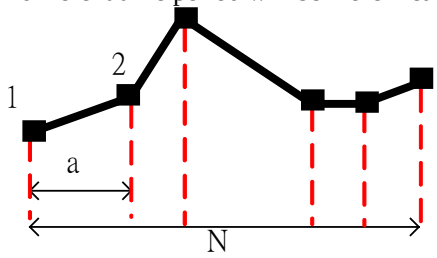

Figure 7. The relative length is *a/N*.

3.5 Matching Analysis

After matching similar curves, the trend in the following n days is analyzed. The future trends of the following *n* days will be concluded by the majority voting method. Ultimately, the resulting pattern is determined if it is significantly meaningful by setting the voting threshold. In Section 4, all samples will be presented as examples.

4. Simulation

Data Resource: Daily stock closing price from Securities & Futures Institute between 1994/1/5 and 2003/7/16.

Experiment 1: Identical industry with the same time interval.

From an industry in Taiwan, Experiment 1 takes Taiwan Semiconductor Manufacturing Company (TSMC) and United Microelectronics Corporation (UMC) as the subjects. The six feature points are extracted for matching within 50 days. The angle error ε is set to be within [-8°, 8°], and the relative length β is equal to 0.1. The result is shown in Figure 8.

Experiment 2: Different industries with the same time interval.

From different industries in Taiwan Experiment 2 takes Taiwan Semiconductor Manufacturing Company (TSMC) and Formosa Plastics Corporation (FPC) as subjects. The six feature points are extracted for matching within 50 days. The angle error ε is set to be within [-8°, 8°], and the relative length β is equal to 0.1. The result is shown in Figure 9.

Experiment 3: Identical industry with different time intervals.

From an industry in Taiwan, Experiment 3 takes Semiconductor Manufacturing Company (TSMC) and United Microelectronics Corporation (UMC) as subjects within different time intervals. The six feature points are extracted for matching within 60 days and 50 days, respectively. The angle error ε is set to be within [-8°, 8°], and the relative length β is equal to 0.1. The result is shown in Figure 10.

Experiment 4: Different industries with different time intervals.

In different industries in Taiwan, Experiment 4 takes Semiconductor Manufacturing Company (TSMC) and Formosa Plastics Corporation (FPC) as subjects within different time intervals. The six feature points are extracted for matching within 60 days and 50 days, respectively. The angle error ε is set to be within [-8°, 8°], and the relative length β is equal to 0.1. The result is shown in Figure 11.

Experiment 5: Self-matching on the same stock.

Do the self-matching on the same stock to search for the repeatedly appearing trend. Experiment 5 takes United Microelectronics Corporation (UMC) and Taiwan Semiconductor Manufacturing Company (TSMC) as subjects. The six feature points are extracted for self-matching within 50 days. The angle error ε is set to be within [-8°, 8°], and the relative length β is equal to 0.1. The result is shown in Figure 12 and 13.

4.1 Simulation Result

In Experiment 1 and 2, it is concluded that the proposed approach in this paper can be applied to either the identical or different industries. In other words, the investment behavior may be the same even between low-correlated industries.

In Experiment 1, TSMC and UMC, both belonging to the electronics industry, have similar curves within [2001/2/21, 2001/5/4] and [2000/11/17, 2001/1/31], as shown in Figure 8 (a1) (b1). It indicates that the strategies of investors are similar because time-space background and influential effects are very close. It is found that the trends of the two stocks are similar in the following 10 days. For example, the trends of the two stocks go downward in the following 10 days, which are shown in Figure 8.

In Experiment 2, TSMC belongs to electronics industry, while FPC belongs to plastics industry. There is low correlation between the two stocks; however, the similar curves appeared in different time interval, as shown in Figure 9. It indicates that the investment strategy may be the same even in different industries.

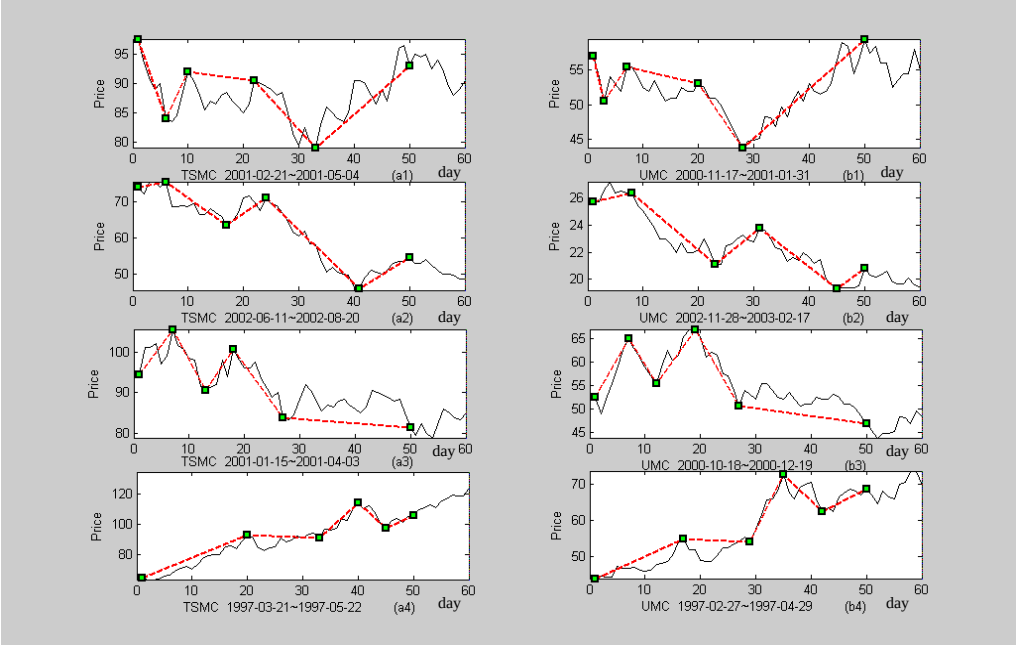

Figure 8. Experiment 1: The figures on the left (TSMC) respectively correspond to the right ones (UMC). (i.e. (a1) is similar with (b1)). The original data is represented by solid curves, the future trend is represented by dotted lines, and the square points stand for feature points. The previous 50 days are for matching, while curves of the last 10 days are the following trend.

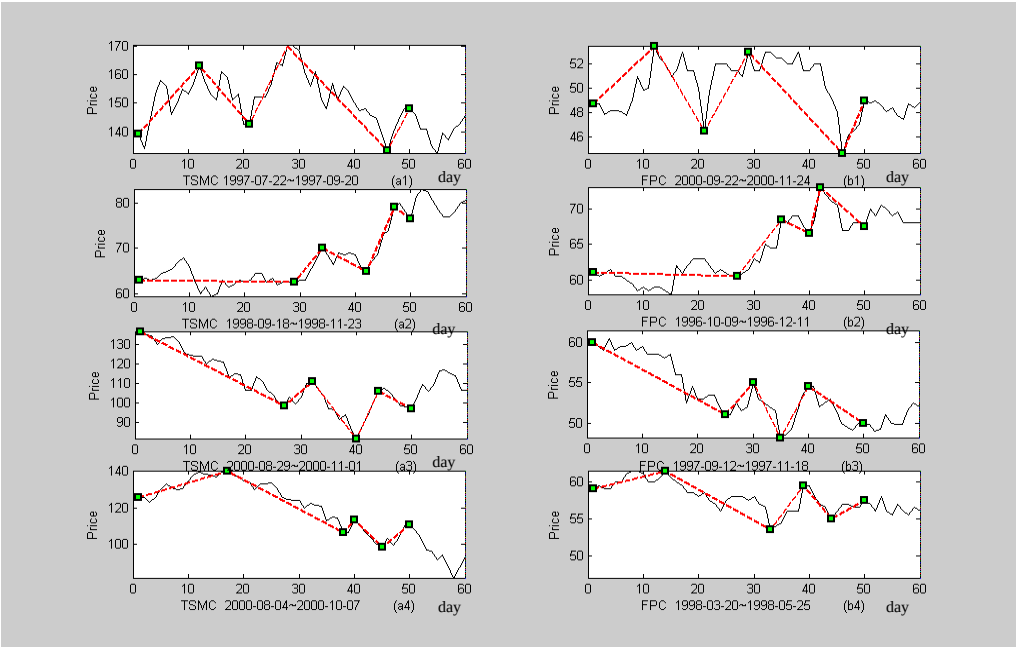

Figure 9. Experiment 2: The figures on the left (TSMC) respectively correspond to the right ones (FPC). (i.e. (a1) is similar with (b1)). The original data is represented by solid curves, the future trend is represented by dotted lines, and the square points stand for feature points. The previous 50 days are for matching, while curves of the last 10 days are the following trend.

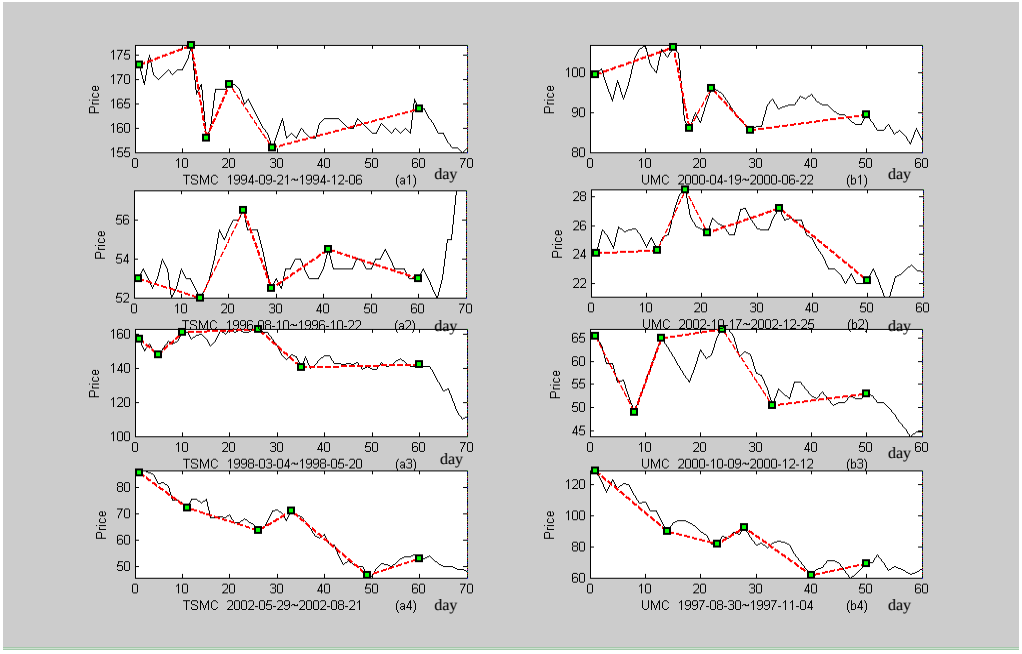

Figure 10. Experiment 3: The figures on the left (TSMC) respectively correspond to the right ones (UMC). (i.e. (a1) is similar with (b1)). The original data is represented by solid curves, the future trend is represented by dotted lines, and the square points stand for feature points. The previous 60 days for the left figures and 50 days of the right ones are for matching, while the curves of the last 10 days are the following trend.

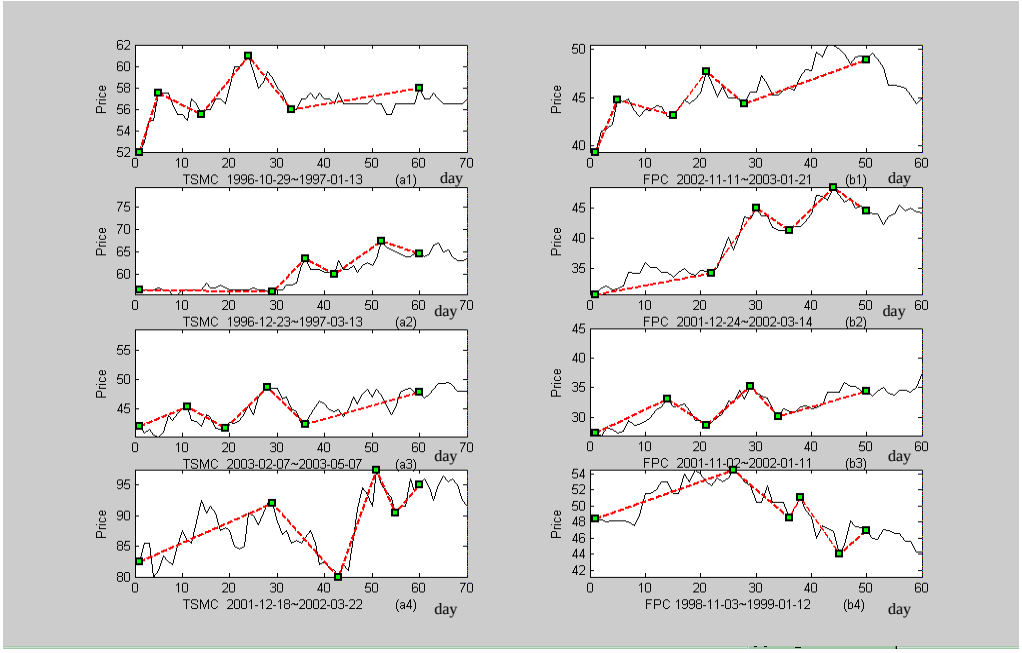

Figure 11. Experiment 4: The figures on the left (TSMC) respectively correspond to the right ones (FPC). (i.e. (a1) is similar with (b1)). The original data is represented by solid curves, the future trend is represented by dotted lines, and the square points stand for feature points. The previous 60 days for the left figures and 50 days of the right ones are for matching, while the curves of the last 10 days are the following trend.

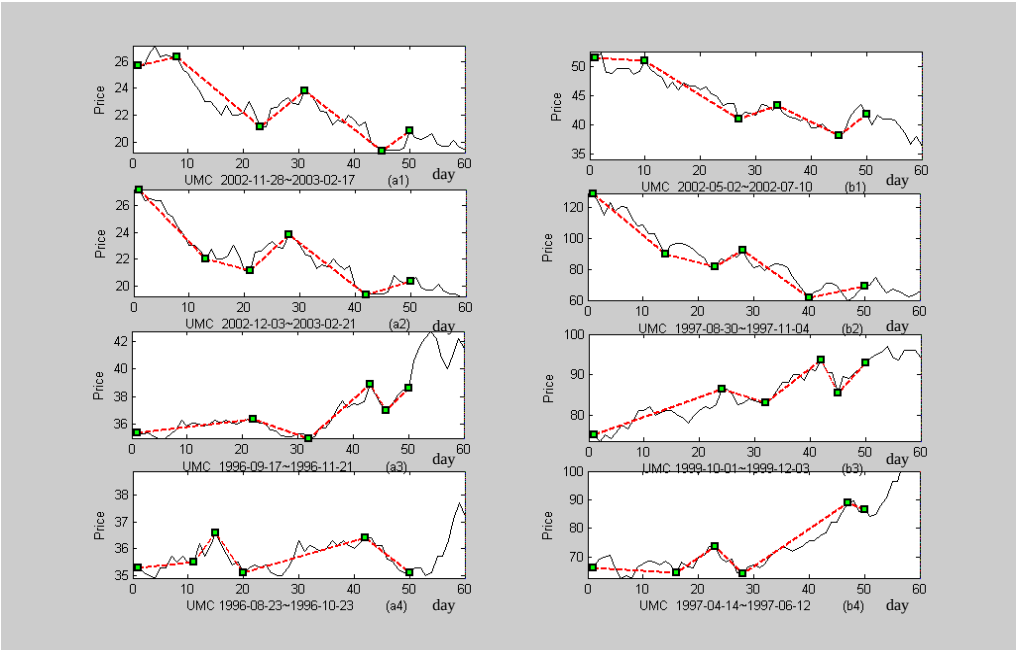

Figure 12. Experiment 5: The figures on the left (UMC) respectively correspond to the right ones (UMC) for self-matching. (i.e. (a1) is similar with (b1)). The original data is represented by solid curves, the future trend is represented by dotted lines, and the square points stand for feature points. The previous 50 days are for matching, while curves of the last 10 days are the following trend.

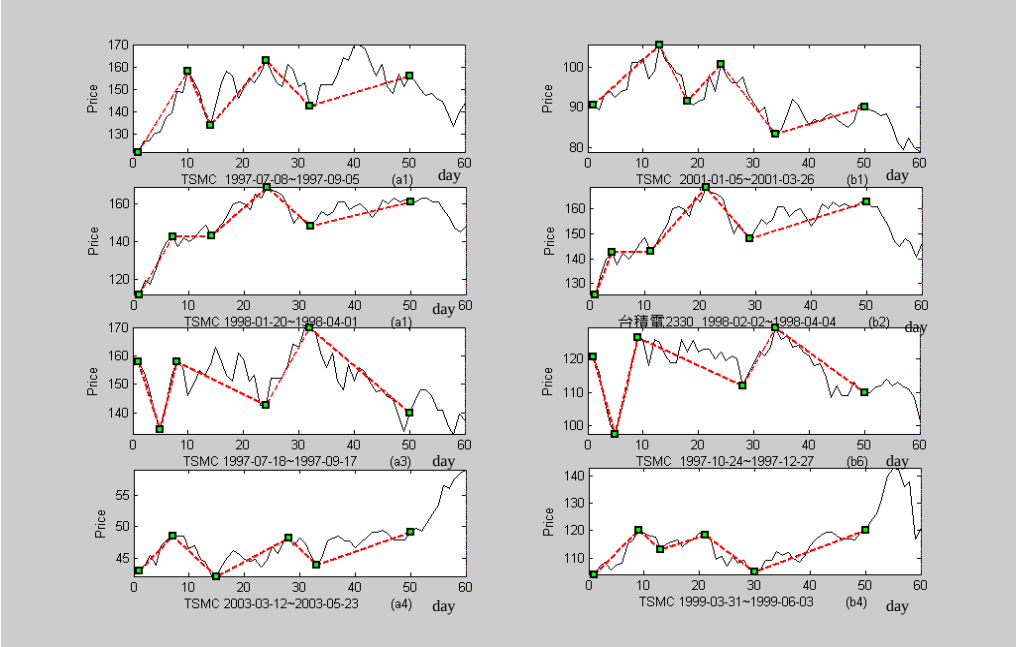

Figure 13. Experiment 5: The figures on the left (TSMC) respectively correspond to the right ones (TSMC) for self-matching. (i.e. (a1) is similar with (b1)). The original data is represented by solid curves, the future trend is represented by dotted lines, and the square points stand for feature points. The previous 50 days are for matching, while curves of the last 10 days are the following trend.

This is because the influential factors may be dissimilar during the same time interval. Nevertheless, the trend in the following 10 days is varied in accordance with different industries.

In Experiment 3 and 4, it is concluded that the proposed approach in this paper can be applied to different time intervals. Because the influential factors will differ according to the stock, the similar curves may appear in different time intervals.

In Experiment 3, different time intervals applied to the same industry can be matched with similar curves. It indicates that the similar trends will still appear in the following 10 days, which is shown in Figure 10 (a3) (b3), and (a4) (b4).

In Experiment 4, different industries with different time intervals can result in the similar curves. The reasons will be (1) all factors that affect the stock price will constitute the identical background even in different industries during different time intervals. (2) The long-term or short-term investment strategies will result in the similar curves. In other words, magnifying a segment of long-term curves is similar with the original long-term curves. However, it is found that in the following 10 days, the trends are not so significantly similar. Perhaps the investors have different judgment standard on different industries.

In Experiment 5, it is found that self-matching on the same stock, as shown in Figure 12 (UMC) and 13 (TSMC) respectively, will match the similar curves. Moreover, the trends in the following 10 days are still similar. In Figure 12 (a4) (b4) and Figure 13 (a2) (b2), the upward trend and downward trend are significantly appeared in the following 10 days respectively.

5 Conclusion

This paper applies feature extraction to stock time series data on critical feature points for similarity matching. The proposed approach is adapted to match either various industries or low-correlated stocks regardless of its data characteristics. The resulting Experiment indicates that stock price volatility will repeatedly appear along with investment strategies. Hence, investors can make current strategies according to the historical repeatedly appearing patterns.

The quantity of feature points is dynamically determined according to stock price data. It is important to describe the original data more appropriate, hasten the searching speed, and improve the accuracy. In addition, the stock volume can also be considered such that the effects will be improved greatly.

References

[1] Adla Jeewook Kim, “Input/output Hidden Markov Models for Modeling Stock Order Flows”, MIT Artificial Intelligence Laboratory, 2001

[2] Chia-Hsien Lin, “The Empirical of Technical Analysis of China A-Share Stock Market”, Master thesis, National Taiwan University, 2002

[3] Fu-Lai Chung, Tak-Chung Fu, Robert Luk, and Vincent Ng, “Evolutionary Time Series Segmentation for Stock Data Mining”, *Data Mining, Proceedings. IEEE International Conference*, 2002, pp83-90.

[4] Guoqing Chen, Qiang Wei, and Hong Zhang, ”Discovering Similar Time-series Patterns with Fuzzy Clustering and DTW Methods”, *IFSA World Congress and 20th NAFIPS International Conference,* 2001, pp.2160- 2164.

[5] Jui-Ching Chen, “On the Prediction of Taiwan Stock Index with Macroeconomic Factor — A Comparative Study of Artificial Neural Network and Multiple Regression”, Master thesis, National Chiao Tung University, 2000.

[6] Mark Last, Yaron Klein, and Abraham Kandel, “Knowledge Discovery in Time Series Databases”, *Systems, Man and Cybernetics, Part B, IEEE Transactions on* ,2001, pp.160-169.

[7] Saswat Anand and Wei-Ngan Chin and Siau-Cheng Khoo, “Charting Patterns on Price History”, *ACM SIGPLAN Notices, Proceedings of the sixth ACM SIGPLAN International Conference on Functional Programming*, 2001, pp. 134-145.

[8] Wen-Po Chung, “Research and Application of Fractal-Dimensional Analysis”, Ph.D. thesis, National Chiao Tung University, 2001.

[9] Yung-Yu Chao, “An Empirical Study on Stock Market Timing with Technical Trading Rules”, Master thesis, National Sun Yat-sen University, 2002.

[10] Zong-Ting Tu, “The Intrinsic Value of MSCI Taiwan Index”, Master thesis, National Chengchi University, 2002.