A Real Time Hybrid Pattern Matching Scheme for Stock Time Series

Zhe Zhang[1], Jian Jiang[2], Xiaoyan Liu[3], Ricky Lau[4], Huaiqing Wang[4], Rui Zhang[3]
[1,4]Department of Information Systems, City University of Hong Kong, Hong Kong
[2]Centre for Computational Finance and Economic Agents, University of Essex, UK
[3]Department of Computer Science and Software Engineering, University of Melbourne, Australia

[1]mailzhezhang@gmail.com, [4]{Ricky.lau,iswang}@cityu.edu.hk
[2]jjiangc@essex.ac.uk
[3]{xiaoyanl,rui}@csse.unimelb.edu.au

Abstract

Pattern matching in stock time series is an active research area in data mining. We propose a new real-time hybrid pattern-matching algorithm in this paper. The algorithm is based on the Spearman's rank correlation, rule sets and sliding window. The concept of sliding windows enables patterns matching to be performed only based on subsequence of stock data which are freshly received. Therefore the proposed algorithm can be applied in real-time application and processing time can be reduced. Spearman's rank correlation coefficient is used to classify the preferred patterns effectively and efficiently first and use the rule sets to provide further ability for describing the query patterns so that is more effective, sensitive and constrainable in distinguishing individual patterns. Encouraging experiment is reported from the tests that the proposed scheme outperforms the other methods both effectively and efficiently, especially in differentiating the special preferred stock patterns or even distorted patterns.

Keywords: Spearman's rank correlation coefficient, Stock time series, Pattern matching.

1 Introduction

The similarity stock pattern search has attracted the attention of the both business and technical experts in recent years. Technical analysts and traders believe that certain stock chart patterns and shapes are signals for profitable trading opportunities. So it is fundamental important to define an effective pattern matching scheme for stock time series.

Much work has been done on performing dimension reduction as the pre-processing step for similarity search to support efficient retrieval and matching of time series. Some of the commonly used methods include Discrete Fourier Transform (DFT) (Agrawal 1993, Rafiei 1997, Faloutsos 1997), Discrete Wavelet Transform (DWT) (Popivanov and Miller 2002, Struzik and Siebes 1999), and Piecewise Aggregate Approximation (PAA) (Yi and Faloutsos 2000).

To measure the similarity between sequence data, several distance metrics are commonly used, such as Euclidean distance (Agrawal 1993), Dynamic Time Warping (DTW) distance (Berndt and Clifford 1994), the slope distance (Toshniwal 2005) between the slopes of the lines approximating the two sequences. We call the two sequences similar if the distance between them is less than a user-defined threshold. Time sequences are usually long, so the distance computation can be time consuming. A solution is to map time sequences into the frequency domain using the Fourier transform, and to use the first few coefficients to filter out non-similar data.

Besides the time consuming problem, another problem with this approach is that the user has no control over the meaning of similarity other than providing a threshold. However, there are some special patterns for stock time series as shown in Figure 1. And it is necessary to recognize and differentiate the patterns considering the amplitude of individual patterns.

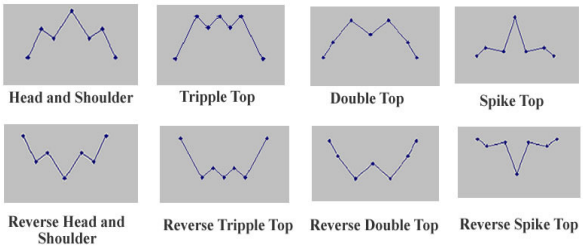


Figure 1: Eight basic patterns for stock data

In this paper, we propose a flexible real time hybrid pattern-matching scheme. The sliding window based principle is involved in pattern matching to reduce the processing time for its online use. And we use Spearman's Rank Correlation Coefficient to classify the preferred patterns effectively and efficiently first, and use the rule sets to provide further ability for describing the query patterns so that is more effective, sensitive and constrainable in distinguishing individual patterns. It outperforms the other methods both effectively and efficiently, especially in differentiating the special preferred stock patterns or even distorted patterns. We will demonstrate the efficiency and effectiveness of the method via extensive experiments based on subsequence matching queries against the real stock price dataset as well as the synthetic dataset.

The remainder of the paper is organized as follows. We introduce the related work in Section 2. In Section 3, we present our proposed method in detail as well as how to use existing techniques. In Section 4, we report experimental results to compare the proposed method and the competitors. Finally, it goes to the conclusion in Section 5.

2 Related Work

There has been much work on similarity-based time series pattern matching. For example, Agrawal et al. (1993) propose an efficient index structure to retrieve time sequences similar to a given one. They map time sequences into the frequency domain by applying the discrete Fourier transform and keep the first few coefficients in the index. Two sequences are considered similar if their Euclidean distance is less than a user-defined threshold.

Besides the DFT, the Discrete Wavelet Transform (DWT) has also been proposed to reduce the number of dimensions of feature vectors in time series (Popivanov and Miller 2002, Shahabi et al 2000, Wu et al 2000). Chan and Fu (1999) use the Haar Wavelet Transform for time series indexing.

While both DFT and DWT focus on mapping the time sequences into other domain, some research focus on processing the time sequences directly in time domain. Many researchers use Piecewise approximation methods as basic to represent the feature extraction (Man and Wong 2001, Morinaka 2001, Yoshikawa 1997). For example, Shatkay (1996) proposes a new notion of generalized approximate queries and proposed a framework that supports them.

Jagadis (1998) uses a fixed gradient alphabet of three letters to represent the patterns, where each letter describes a movement of direction and covers a specific length of time. And Toshiniwal and Joshi (2005) consider the time series as a pattern defined by a set of regular expressions of slopes. This technique is based on the intuition that similar time sequences will have similar variations in their slopes and consequently in their time weighted slopes.

For the similarity pattern search method, the Euclidean distance is the mostly widely utilized metric to measure the similarity between the query and candidate sequence (Toshiniwal and Joshi 2005, Keogh and Pazzani 2000, Agrawal 1993, Goldin and Kanellakis 1995, Rafiei and Mendelzon 1997). According to this method, if the Euclidean distance $D(X, Y)$ between two time sequences X and Y of length n is less than a threshold å, then the two sequences are said to be similar. The Euclidean distance is given as:

$$D = (X, Y) = \sqrt{\sum_{i=1}^{n} (X_{i-} Y_i)^2} \qquad (1)$$

A major shortcoming in the Euclidean distance is that it only considers the whole shifts of two sequences and is not able to handle vertical and horizontal shifts separately, which exist between the time sequences under comparison.

Toshniwal et al (2005) have used the cumulative variation of slopes for computing the similarity in the given time series data, which take into account of the ratio of vertical and horizontal shift. In this technique, the X_i, Y_i were substituted by the slopes Sc_j and Sq_j, which were for the jth strip in the candidate time sequence C and the query time sequence Q respectively. And Toshniwal et al (2006) also presented an approach for similarity search in time series data, which is an improvement over the former one, and add time-weighted coefficient on the slope.

However, there are some special requirements for the stock data pattern matching due to their specialized shape for frequently occurred patterns as shown in Figure 1. Therefore, it is required to consider the separate shift for each point in the patterns found in stock data for differentiating the specific patterns of stock when doing pattern similarity comparison. On the other hand, most of the patterns are distorted from the basic patterns (Fu et al, 2007), as shown in Figure 2, but they still should be retrieved during pattern searching. Therefore, we propose the new pattern-matching scheme, which can recognize and differentiate the special preferred stock patterns or even distorted patterns and make a faster and more accurate matching. In the following section, we will propose our hybrid pattern matching approach.

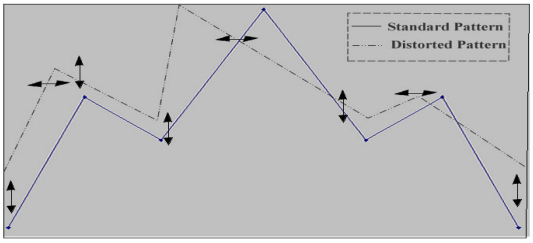


Figure 2: Distorted pattern

3 Stock Pattern Matching Based on Hybrid Pattern Matching Approach

Indeed, there are two main problems in pattern matching: how to define the preferred patterns for query and how to match the patterns with the pattern template in different resolution (Fu et al, 2007). We propose a flexible online pattern-matching scheme based on sliding window, which is involved in the whole matching process, including both in feature point extraction and pattern matching. We modify the feature point extraction method based on Perceptually Important Point (PIP) into sliding window based and being determined by the feed back of pattern matching outcome, fulfilling the demand of time saving and online use. On the first step, we propose a pattern-matching scheme, which based on Spearman's Rank Correlation Coefficient to classify the preferred patterns effectively and efficiently. And secondly we use the rule sets to provide further ability for describing the query patterns and is more effective, sensitive and constrainable in distinguishing individual patterns.

3.1 Pattern matching based on sliding window

A good representation or approximation of time series is important for time and memory efficiency of the searching algorithms (Man and Wong 2001). Time series pattern matching based on Perceptually Important Point (PIP) identification is introduced by Chung et al. (2007), which is used for feature point extraction. The principle of PIP algorithm is to capture the fluctuation of the sequence and take these highly fluctuated points as PIPs. Firstly, the first two PIPs are defined as the first and last point of input sequence P. The next PIP will be the point in P with maximum distance D to the first two PIPs. The fourth PIP will be the point in P with maximum D to its two adjacent PIPs. The process continues until the length of extracted

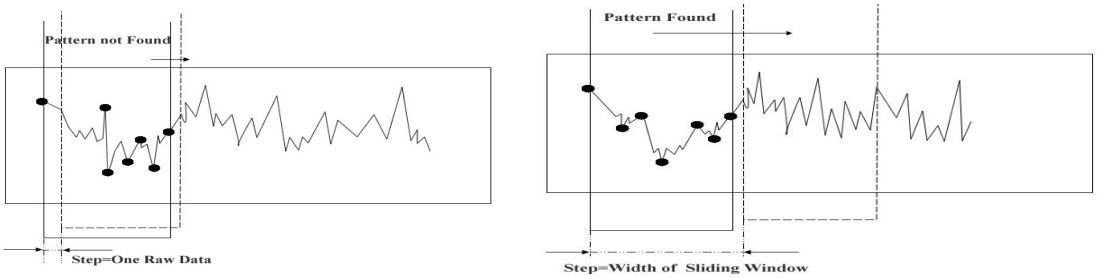


Figure 3: Process for sliding window movement

series (SP) is equal to that of query sequence Q.

The PIP based algorithm is appropriate to be used for stock data (Zhang, et al 2006; Jiang et al, 2007). And Zhang et al (2007) modified the maximum distance D into another coefficient to make it appropriate to stock characteristic. Chung et al. (2007) also modified the distance D into different forms and found that it was efficient and effective to extract the feature points when D was perpendicular distance between the test point and the line connecting the two adjacent PIPs. Therefore, PD was used in PIP calculation in the proposed algorithm.

We define a window width W for sliding window and put the PIP-PD based algorithm processed in each sliding window. The window moves step by step for searching the patterns. The step length of sliding window is determined by the pattern matching outcomes. There are two situations: (1) If the preferred pattern is found, the window will move the length, which is equal to the window; (2) if no preferred pattern is found, the window will move just one raw data to do pattern matching until the preferred pattern is found and it goes to situation (1). Figure 3 shows the process for sliding window movement. And Figure 4 shows the Sub procedure 1 for sliding window PIP-PD.

In such way, we can accelerate the pattern matching and make sure most of the patterns will not be skipped. And the sliding window scheme makes the algorithm available to online use while the old ones cannot. The main pseudo code for the pattern matching is shown in Figure 5. The detailed processes of Pattern Matching Procedure to sliding window series (Sub procedure 2) are shown in the next section.

Sub Procedure 1: Apply sliding windows on P[1:n] to extract sliding window-length feature point, within sliding window, SW[1:W]

Input: Raw_Dada(P[1:n]) for each window
Output: SP[1:N]
PIP procedure
  Set SP[1]=SW[1]
  Set SP[N]=SW[W]
  Do
    Select SW[i] with maximum PD to the adjacent points in SW
    SP[j]=SW[i]
  Until j=N
End PIP

Figure 4: Sub procedure 1 for sliding window PIP-PD

3.2 Pattern matching based on Spearman's rank correlation coefficient

In this section, we will describe the sub procedure 2 in detail. Technical analysts and traders believe that certain stock chart patterns and shapes are signals for profitable trading opportunities. Many professional and amateur traders claim that they consistently make trading profits by following those signals. We propose eight basic patterns (Figure 1) which are proved to be useful in stock trading and defined by LO (2000), in ranking format based on Spearman's rank method.

Main Procedure: Sliding Window Pattern Matching

Input: full time series data, RAW_DATA (P [1:n])
Initialize: Set Window Width, W
  Set the number of feature points extracted from time series within sliding window, N
  Set Pattern Template PAT_TEMP [1:8]
Output: Confirmed Pattern

DO
  Sub procedure 1: Apply sliding windows on P[1:n] to extract sliding window-length feature point, within sliding window, SW[1:W]
  End of Sub procedure 1
  Sub procedure 2: Apply Pattern Matching Procedure to SP[1:N] to get matched patterns from PAT_TEMP[1:8], the matched pattern is PAT
  End of Sub procedure 2
    If PAT exists
      Move the window with step= W
    Else
      Move the window with step= next raw data – SW[1]
    End if
Until finish the length of P

Figure 5: Main pseudo code for the pattern matching

The Spearman rank correlation coefficient is a nonparametric technique for evaluating the degree of correlation between two variables. It operates on the ranks of the data rather than the raw data. There are some advantages for using Spearman's rank correlation coefficient over the more common product moment correlation coefficient to define the stock pattern templates.

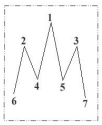
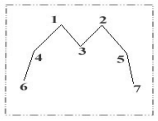
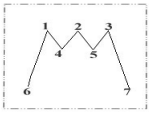
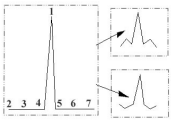
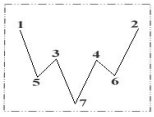
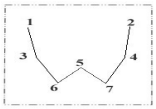
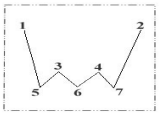
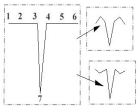
Patterns Name	Head & Shoulder	Double Top	Triple Top	Spike Top
Pattern Figure				
Position In the Descending Order	[6 2 4 1 5 3 7]	[6 4 1 3 2 5 7]	[6 1 4 2 5 3 7]	[2 3 4 1 5 6 7]
Rank	[6.5 2.5 4.5 1 4.5 2.5 6.5]	[6.5 4.5 1.5 3 1.5 4.5 6.5]	[6.5 2 4.5 2 4.5 2 6.5]	[4.5 4.5 4.5 1 4.5 4.5 4.5]
Patterns Name	Head & Shoulder (Reversed)	Double Top (Reversed)	Triple Top (Reversed)	Spike Top (Reversed)
Pattern Figure				
Position In the Descending Order	[1 5 3 7 4 6 2]	[1 3 6 5 7 4 2]	[1 5 3 6 4 7 2]	[1 2 3 7 4 5 6]
Rank	[1.5 5.5 3.5 7 3.5 5.5 1.5]	[1.5 3.5 6.5 5 6.5 3.5 1.5]	[1.5 6 3.5 6 3.5 6 1.5]	[3.5 3.5 3.5 7 3.5 3.5 3.5]

Table 1: Eight model patterns and ranks

There are several advantages for using Spearman's rank to define the stock pattern templates.

(1) Only one sorting process is needed to treat the seven different patterns. It accelerates the speed for pattern patching while other methods need to apply different approaches to different patterns.

(2) It can recognize more distorted patterns because it only considers the rank instead of the real data for each point. So it tolerance s a shift range for each point when it ranks in the same position. For example, for the Head & Shoulder pattern, there is no restrict for whether the left shoulder should be higher than the right shoulder or not. So this method can achieve the requirement because the assigned rank number is equal for the two shoulders in our rank scheme (e.g. both are 2.5). For the Spike Top pattern, there are two situations just as shown in the pattern figure, so we define the pattern template in such way to make the pattern generalized.

(3) It operates on the ranks of the data it is relatively insensitive to outliers and there is no requirement that the data be collected over regularly spaced intervals. (Gauthier, 2001).

The raw scores are converted to ranks, and the differences ñ between the ranks of each observation on the two variables are calculated. We use the Eq. (2) to calculate the Spearman's Correlation Coefficient because there is no large number of tied ranks (Gauthier 2001). Each of the points with equal value should be assigned same rank. It is an average of their positions in the ascending order of the values. Then ñ is given by:

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \qquad (2)$$

where d_i is the difference between ranks for each x_i, y_i data pair, and n is number of data pairs.

The steps for our pattern matching scheme in sub procedure 2 are: 1) Assign the rank to the feature points extracted from sub procedure 1, get SP_R[1:N]. 2) Calculate SP_R[1:N] with the template, PAT_TEMP[1:8] to get the spearman coefficient, SP_C[1:8]. 3) Comparing the SP_C[1:8] with threshold of each template, THRE[1:8] and get the post pattern, POS_PAT. 4) Evaluate whether the POS_PAT can pass the rules defined for each pattern. 5) Put those POS_PAT that pass the rules and with maximum Spearman's coefficient in MATCH_PAT. The sub procedure 2 is shown the Figure 6.

Sub Procedure 2
Input: SP, PAT_TEMP[1:8]
Output: MATCH_PAT
Assign the Rank to SP[1:N], SP_R[1:N]
Compare the SP_R with the Ranks of each PAT_TEMP [1:8] to get Spearman Coefficient, SP_C[1:8]
If SP_C [i] > Threshold of each template, THRE [i] (i = 1: 8)
Include PAT_TEMP[i] in the array of Possible patterns POS_PAT
End if
Sub Procedure 2.1:
Check whether POS_PAT can pass the defined rules for each template.
End Sub Procedure 2.1
Add POS_PAT which can pass the RULES for each predefined templates of patterns AND maximum Spearman coefficient to Match_Pat
Return MATCH_PAT
End of Sub Procedure 2

Figure 6: Pseudo of sub procedure 2

The disadvantage of this technology is that there is a loss of information when the data are converted to ranks. So we defined a set of rules for each pattern to provide further ability for describing the query patterns. The rule-based method can overcome the shortcoming of spearman's correlation coefficient method for it can recognize the specific patterns more explicitly. For example, in double top pattern, the two top's amplitudes should be within 15% difference according the definition of Lo et al. (2000).

The rule sets for sub procedure 2.1 is shown in Figure 7. And a set of user-defined parameters has to be set as follows.

Max1:	The maximum number of seven extracted points, SP[1:7] in sliding window
Max2:	The second maximum number of seven extracted points, SP[1:7] in sliding window
Min1:	The minimum number of seven extracted points, SP[1:7] in sliding window
Min2:	The second minimum number of seven extracted points, SP[1:7] in sliding window
SP[i]:	The ith position in seven points from the left to the right
%:	The percentage of difference between highest and lowest point

Time Scaling

Change the sequence length of the pattern templates from P[1…n] to P[1… m], where n=7 and m=25,43 and 61 (which means 4 , 7 and 10 times length of original templates)

Time Warping

For each critical point p[i] in P

Move p[i] between p[i–1] and p[i+1] randomly with width W

$W \in [t_{p[i]}-(t_{p[i]}-t_{p[i-1]})/3],\ t_{p[i]}+(t_{p[i+1]}-t_{p[i]})/3]$ or

$W \in [t_{p[i]}-(t_{p[i]}-t_{p[i-1]})*2/3,\ t_{p[i]}-(t_{p[i]}-t_{p[i-1]})/3 \cup [t_{p[i]}+(t_{p[i+1]}-t_{p[i]})/3,\ t_{p[i]}+(t_{p[i+1]}-t_{p[i]})*2/3]$ or

$W \in [t_{p[i-1]},\ t_{p[i]}-(t_{p[i]}-t_{p[i-1]})*2/3] \cup [t_{p[i]}+(t_{p[i+1]}-t_{p[i]})*2/3,\ t_{p[i+1]}]$

End for

Noise Adding

For each data point p[i] in P

If randomly generated probability < prob (supposed to be 0.5)

diff=(p[i+1]-p[i]) * random_amplitude between 0 to ampl

ampl∈[0,0.1] or

ampl∈[0.1,0.2] or

ampl∈[0.2,0.3]

p[i]=p[i]±diff

End if

Figure 8: Psedo code of generating synthetic pattern templates

4 EXPERIMENT

We evaluate our proposed method in this section. The proposed method is compared with two competitors: Euclidean distance based method and slope based method, which are popular and introduced in section 2. The accuracy is used to compare the three methods. It is defined as the percentage of the number of correctly matched patterns when a query pattern is given, and calculated as follows. We run experiments on both synthetic sequences and real stock price data, the past 21 years Hong Kong Hang Seng Index (5087 data points). And we use the definition proposed by Lo (2000) as the standard pattern:

$$\text{Accuracy} = \frac{\text{Number of correctly matched patterns}}{\text{Number of total patterns}} \quad (3)$$

Rule sets for eight patterns in sub procedure 2.1

Rule Set for Double Top Pattern, Orientation= Up (Down)

Rule 1: Max1 - Max 2 < 15%
(Rule 1' for Down: Min1 - Min 2 < 15%)
Rule 2: the two maximum (Rule 2': minimum) points are the 3rd and 5th point
(Rule 2' for Down: the two minimum points are the 3rd and 5th point)
Rule 3: Rank of SP2>Rank of Sp1 and rank of Sp7<Rank of Sp6
(Rule 3' for Down: Rank of SP2<Rank of Sp1 and rank of Sp7>Rank of Sp6)

Rule Set for Head & Shoulders Pattern, Orientation= Up (Down)

Rule 1: |SP2 – SP6| < 15%
Rule 2: |SP3 - SP5| < 15%
Rule 3: the ranking of SP4 is first
(Rule 3' for Down: the ranking of SP4 is last)
Rule 4: the ranking of sp2 and sp6 must be 2 and 3
(Rule 4' for Down : the ranking of sp2 and sp6 must be 5 and 6)
Rule 5: the ranking of sp1 and sp7 must be 5 or 6 or 7

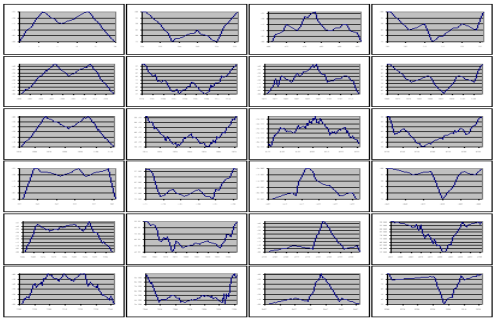
Rule Set for Triple_tops Pattern, Orientation= Up (Down)

Rule 1: Max(|SP2 - SP4|, |SP2 - SP6|, |SP4 - SP6|) < 15% % used to separate from Head and shoulders
Rule 2: |SP3 - SP5| < 15%
Rule 3:sp2, sp4, sp6 must be 3 highest points
(Rule 3' for Down: sp2, sp4, sp6 must be 3 lowest points)

Rule Set for Spike_top Pattern, Orientation= Up (Down)

Rule 1: time_sequence of Max1 = 4
(Rule 1' for Down: time_sequence of Min1 = 4)
Rule 2: SP4 - MAX2 >=75%
% used to separate from Head & Shoulders
(Rule 2' for Down: SP4 - MIN2 >=75%

Figure 7 Rule-based sets for sub procedure 2.1



Double Top	Reversed Double Top	Head & Shoulder	Reversed Head & Shoulder
Triple Top	Reversed Triple Top	Spike Top	Reversed Spike Top

Figure 9: Sample of synthetic sequence for eight patterns

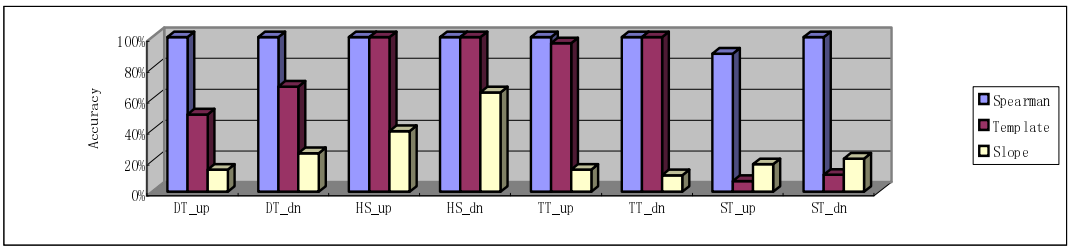


Figure 10: Accuracy for three methods

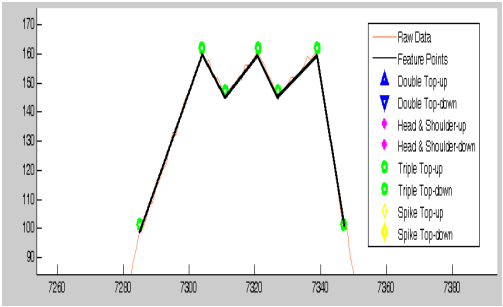


Figure 11: Triple pattern is found by our method

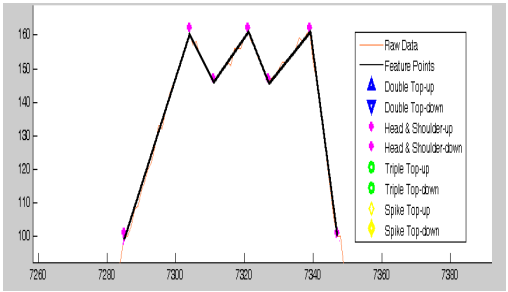


Figure 12: Slope method mistakes triple top pattern for head and shoulder pattern

4.1 Synthetic Data

We generate 216 synthetic sequences for 8 patterns shown in Figure 1. There are three parameters for generating the sequences by improving more details of the parameters than Fu et al (2007) proposed. The *time scaling* is used to scale the length of time series, and is defined for 3 levels: short scaling, middle scaling and long scaling. The *time warping* is used to change the position of critical points with three 3 levels: short width warping, middle width warping and large width warping. The noise is used to add three levels amplitudes of noise: small amplitude, middle amplitude, and large amplitude. The Psedo of generating the synthetic sequences is shown in Figure 8.

And we select some sample sequences for eight patterns with three parameters ranged between three levels in Figure 9.

We compare the accuracy of three methods used on the synthetic sequences. Figure 10 shows the accuracy for three different methods applied on eight patterns. Our approach outperforms the other approaches for all of the eight patterns. Because our approach pays more attention on the specific shape of each pattern rather than only comparing the point-distance in the space just like the Euclidian distance method. And for the slope-based method, whose computation is derived from the Euclidian distance, considers the ratio of vertical and horizontal distance for each strip and pays less attention on the length of each strip. However, the specific rules are added in our approach for more detailed shape recognizing for each pattern, so it outperforms the slope-based method.

On the other hand, we can find from Figure 10 that our approach outperforms others especially for Spike Top and Triple Top pattern. Because these two patterns are similar in the shape and the only difference among them is the amplitude of each critical point.

For the Euclidian and Slope-based methods, they are easy to mistake the Triple Top pattern for Head and Shoulder pattern or for other. For example, the pattern shown in Figure 11 is Triple Top. However, the Slope Method mistakes it for Head and Shoulder in Figure 12. So their accuracies for Triple Top are lower than our method.

4.2 Real Data

We apply our approach together with the two competitors on the past 21 years Hong Kong Hang Seng Index (5087

data points).

Firstly, we compare the processing time for three methods with different width of sliding window from small length, middle length and large length. The result is shown in Figure 13. We can find from Figure 13 that the processing time for our method is steady regardless the width of sliding window and approximately the same as the Euclidean Distance method. The processing time is less than the slope method and slightly more than the Euclidian method when the length of sliding window is small. And it outperforms the Euclidian Distance method when the width of sliding window is large, and slightly takes more time than Slope method does. Because the advantage of our approach is that it pre-classifies the patterns due to the Spearman's Rank Correlation Coefficient and then specify the more detailed pattern based on rule sets. So if no pattern matches the Spearman's Coefficient in the first stage, the second stage for verify the detailed pattern is omitted and then the time is saved. So it will outperform the Euclidian Distance method when less patterns found in the roughly pattern shape determination (e.g. when using long sliding window length). When there are more patterns pass the first stage, it should only pay the additional time for the second stage so it needs slightly more time than the Euclidian Distance method. E.g. there are more patterns found when the small sliding window length is used, so our approach costs slightly more time than the Euclidian Distance method.

Another advantage of saving time for our approach is that in the second stage of our method, it reuses the parameters that results from the sorting process in the first stage such as the MAX, MIN. So there is no need to calculate other parameters again in order to save time.

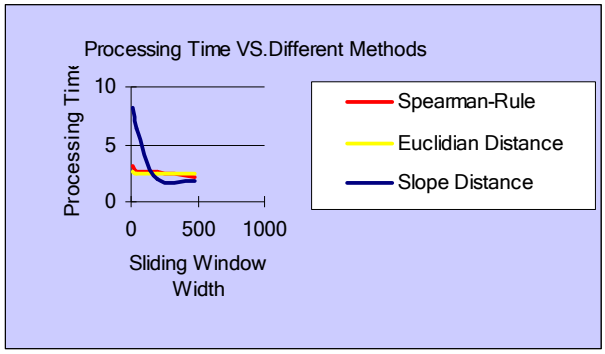


Figure 13: Processing time for three patterns matching methods

Secondly, we compare the performances of different methods for accuracy of identifying eight patterns by changing the length of sliding window. We select the sample when length of sliding window is 40, which is shown in Figure 14. And the zoom-in picture for circled area is shown in Figure 15. We can find from Figure 14 and Figure 15 that our method can recognize the eight patterns effectively.

The whole accuracy for eight patterns using three methods is shown in Figure 16. The result is similar as what is found in synthetic data. And the accuracy of spike top patterns was higher than the synthetic data because there were very little spike top patterns found in real data.

Out method outperforms the other two for each pattern. The accuracy for Head & Shoulder pattern, Triple Top Pattern by using Euclidean Distance method and Slope method are relatively low. Because the Euclidean Distance method only pays attention on the accumulated point shifts and neglects the individual shift of each point, it can not differentiate three similar patterns such as Head & Shoulder, Spike Top Pattern and Triple Top pattern when their only difference is the amplitude fluctuates between the middle point and the two side points. Neither does Slope method. Because although Slope method is a very efficient and effective method and does very well for the general time series for in time series pattern matching, it may not specialized and suitable for the stock pattern matching for stock patterns' characteristic. Therefore, it still can not differentiate the similar patterns as used in stock patterns. However, our method can fulfill the above requirement, by paying attention on the rank of each point and using rules to specify the patterns. Figure 17 shows the Head and Shoulder Pattern is correctly recognized by using our method when the Euclidean Distance method mistakes it for Triple Top pattern in Figure 18.

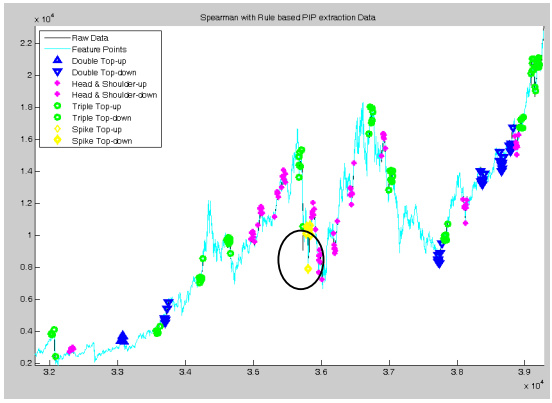


Figure 14: Patterns found using our method (SW-Len=40)

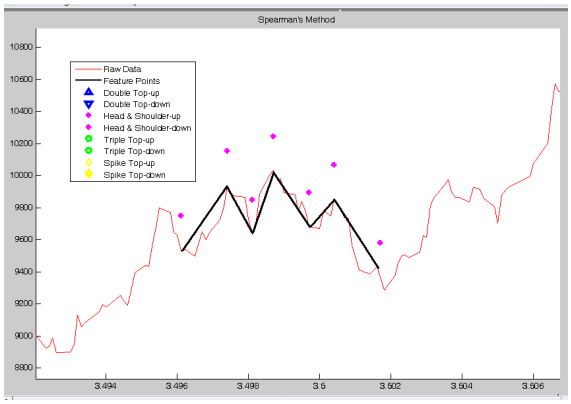


Figure 15: Zoom-in area for recognized head and shoulder pattern

5 Conclusions

We propose a flexible online hybrid pattern-matching scheme, which is a combination of Spearman's Rank Correlation Coefficient and rule sets. We use Spearman's rank correlation coefficient to classify the preferred

patterns effectively and efficiently first and use the rule sets to provide further ability for describing the query patterns so that is more effective, sensitive and constrainable in distinguishing individual patterns. It is much efficient and effective than the current pattern matching approaches such as Euclidian Distance based and Slope based method. The proposed scheme can be used real time for its sliding window based pattern matching and be with less processing time. It can also recognize more distorted patterns with noise. And it outperforms other methods when differentiating the special patterns for stock time series.

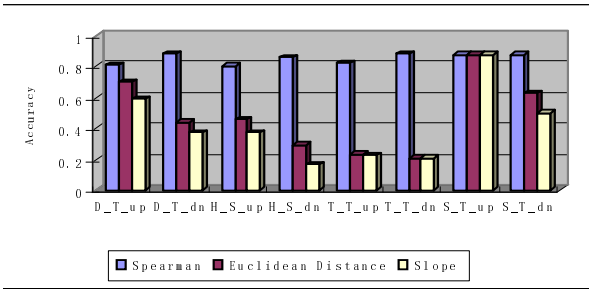


Figure 16: Accuracy for finding eight patterns using three methods

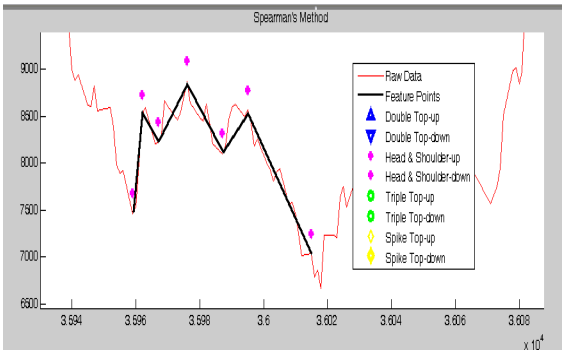


Figure 17: Head and shoulder pattern recognized by using our method

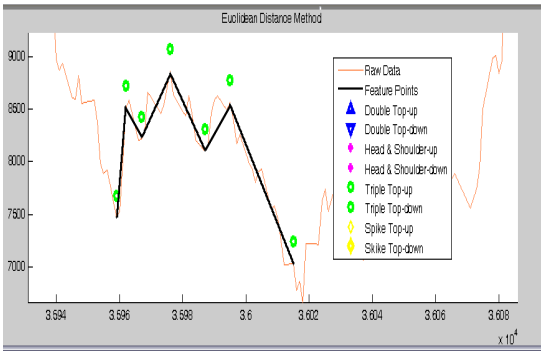


Figure 18: Euclidean distance method mistakes head and shoulder pattern for triple top pattern

Acknowledgement

This work has been supported by the Australian Research Council (ARC) under project DP0880250.

References

Agrawal, R., Faloutsos, C. and Swami, A. (1993): Efficient similarity search in sequence databases. *Proc. of The 4th Int'l Conf. On Foundations of Data Organization and Algorithms*, 69-84.

Berndt, D. J. and Clifford, J. (1994): Using dynamic time warping to find patterns in time series. *Proc. of KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, Seattle, Washington, 359-370.

Chan, K. *and* Fu, A. W. (1999): Efficient time series matching by wavelets. *Proc. of 15th Int'l Conf. on Data Engineering*, 126-133.

Faloutsos, C., Jagadish, H., Mendelzon, A. and Milo, T. (1997): A signature technique for similarity based queries. *Proc. International Conference on Compression and Complexity of Sequences*, Positano-Salerno, Italy.

Fu, T. C., Chung, F. L., Luk, R., and Ng, C. (2007): Stock time series pattern matching: template-based vs. rule-based approaches, *Engineering Application of Artificial Intelligence*, 20(3): 347-364

Goldin, D. Q. and Kanellakis, P. C. (1995): On similarity queries for time-series data: constraint specification and implementation. *Proc. of Constraint Programming 95*, Marseilles, France.

Jiang, J., Zhang, Z., and Wang, H. (2007): A new approach to improve multi-dimensional stock data Reduction, *Proc. of Iadis European Conference on Data Mining*, Lisbon, Portugal.

Keogh, E. and Pazzani, M. (2000): A simple dimensionality reduction technique for fast similarity search in large time series databases. *Proc. of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining* , 122-133.

Keogh, E. and Smyth, P. (1997): A probabilistic approach to fast pattern matching in time series databases. *Proc. of 3rd International Conference on Knowledge Discovery and Data Mining*, 24-20.

Lo, A.W., Mamaysky, H. and Wang, J. (2000): Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. *Journal of Finance* 55 (4), 1705-1765.

Man, P. W. P. and Wong, M. H. (2001): Efficient and robust feature extraction and pattern matching of time series by a lattice structure. *Proc. of 10th International Conference on Information and Knowledge Management*, 271-278.

Morinaka, Y., Yoshikawa, M., Amagasa, T. and Uemura, S. (2001): The l-index: an indexing structure for efficient subsequence matching in time sequence databases. *Proc. of Pacific-Asian Conference on Knowledge Discovery and Data Mining,* 51-60.

Popivanov, I. and Miller, R. J. (2002): Efficient similarity queries over time series data using wavelets. *Proc. of*

The 18th International Conference on Data Engineering, 273-282.

Rafiei, D. and Mendelzon, A. (1997): Similarity -based queries for time series data. *Proc. of the ACM SIGMOD Conference*, Tucson, Az, May.

Shahabi, C., Tian, X. and Zhao, W. (2000): Tsa-Tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries. *Proc. of 12th International Conference on Scientific and Statistical Database Management*, 55-68.

Shatkay, H. and Zdonik, S. (1996): Approximate queries and representations for large data sequences. *Proc. of the 12th International Conference on Data Engineering*, 536-545.

Struzik, Z. and Siebes, A. (1999): The Haar wavelet transform in the time series similarity paradigm. *Proc. Principles of Data Mining and Knowledge Discovery, 3rd European Conference*, Prague, Czech Republic, 12-22.

Thomas, D.G. (2001): Detecting trends using spearman's rank correlation coefficient, *Environmental Forensics*, 359-362.

Toshniwal, D. and Joshi, R. C. (2005): Similarity search in time series data using time weighted slopes, *Informatica* 29(1): 79-88.

Wu, Y., Agrawal, D. and Abbadi, A. (2000): A comparison of DFT and DWT based similarity search in time-series databases. *Proc. of the 9th International Conference on Information and Knowledge Management*, 488-495.

Yi, B. K. and Faloutsos, C. (2000): Fast time sequence indexing for arbitrary LP norms. *The VLDB Journal*, 385-394.

Yi, B. K., Jagadis, H., and Faloutsos, C. (1998): Supporting fast search in time series for movement patterns in multiple scale. *Proc. of the 7th ACM International Conference on Information and Knowledge Management*, 251-258.

Zhang, Z., Liu, X. Y. and Wang, H. Q. (2006): Discovery in stock time series based on perceptually important point algorithm, *Proc. of San Diego International Systems Conference*, San Diego, La, USA.