Applied Soft Computing 38 (2016) 346–359

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Effect of segmentation on financial time series pattern matching

Yuqing Wan, Xueyuan Gong, Yain-Whar Si*

Department of Computer and Information Science, University of Macau, Macau

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form
25 September 2015
Accepted 6 October 2015
Available online 21 October 2015

Keywords:
Segmentation
Pattern matching
Financial time series
Technical pattern
Head-and-shoulders

ABSTRACT

In financial time series pattern matching, segmentation is often performed as a pre-processing step to reduce the data points from the input sequence. The segmentation process extracts important data points and produces a time series with reduced data points. In this paper, we evaluate the effectiveness and accuracy of four approaches to financial time series pattern matching when used with four segmentation methods, the perceptually important points, piecewise aggregate approximation, piecewise linear approximation and turning points methods. The pattern matching approaches analysed in this paper include the template-based, rule-based, hybrid, decision tree, and Symbolic Aggregate approXimation (SAX) approaches. The analysis is performed twice, on a real data set (of Hang Seng Index prices from the Hong Kong stock market) and on a synthetic data set containing positive and negative cases of a technical pattern known as head-and-shoulders.

1. Introduction

In financial trading, two types of analysis are usually used to predict future price movements. The first type, called "technical analysis", involves forecasting a trend in the financial market based on historical data such as the daily prices and volumes traded. The historical data are commonly represented in the form of a time series. The second type is called "fundamental analysis", which is the prediction of price movement based on the financial strength (health) of the company or on developments in the social, political and economic situation. In technical analysis, one of the crucial steps performed by traders is locating interesting patterns in the time series that can help to forecast future price trends. These patterns are commonly referred to as technical patterns (or chart patterns) in the financial domain. Some of the well-known technical patterns include head-and-shoulders (H&S), double top, triple top, cup-with-handle and range breakout [1].

Pattern matching is one of the most important tasks for the analysis of financial time series and many novel pattern matching approaches have appeared in recent years. Zapranis and Tsinaslanidis [2] proposed a new approach based on neural networks to identify a technical pattern known as H&S. Zhou et al. [3] proposed a geometrical similarity measure approach that is invariant to shifting and scaling, and which calculates the angle between two vectors after shift-eliminated transformation. In [4], a kernel regression estimator of a given time series is constructed. The extrema on the original time series is identified based on the local minimum or maximum in the regression line. The extremas in the original time series are then used to determine whether or not a pattern has occurred. Rao and Principe [5] proposed a generalized eigendecomposition algorithm using two step Principal Component Analysis (PCA) process for segmenting speech signals. In [6], Ge et al. proposed an approach to model time series with Hidden semi-Markov Model (HSMM) to detect specific waveform patterns. Symbolic Aggregate approXimation (SAX) [7] allows a time series of arbitrary length to be reduced to a string of arbitrary length. Barnaghi et al. [8] proposed an enhanced SAX which uses K-means clustering method to determine the zones of the symbols. To calculate the similarity measure, Damerau–Levenshtein distance [9,10] is also used to find the edit distance between two strings. Kullback–Leibler divergence [11], Jensen–Shannon divergence [12] or Bhattacharyya distance [13] are also used to compare the difference between two probability distributions.

A number of comprehensive surveys on time series pattern matching are also reported in literature. Fu [14] gives a review on time series data mining and categorizes the time series data mining research into representation, indexing, similarity measure, segmentation and visualization. Xing et al. [15] survey sequence classification methods in terms of methodologies and application domains. A comprehensive survey of control-chart pattern-recognition methods is also reported in [16].

Dynamic time warping (DTW) is one of the most popular similarity measure based approaches for time series pattern matching [17]. Some researchers have applied an extended version of dynamic time warping (DTW) for pattern matching. For instance, Li et al. [18] proposed a novel similarity measure approach based on piecewise linear approximation and derivative DTW. Junkui and Yuanzhen [19] accelerated the DTW process by terminating the calculation earlier when the values of the neighbour cells in the cumulative distance matrix exceeded the tolerance level. Chen et al. [20] proposed a new distance calculation method called DTW-D that combines DTW and Euclidean distance (ED) for time series semi-supervised learning algorithms. In [21], string kernels are used to measure similarity of strings in linear time by using annotated suffix trees. However, computing string kernels using suffix trees does not scale well to problems with large data size [22]. To alleviate this problem, Teo et al. [22] compute string kernels by designing a space efficient and scalable algorithm using Enhanced Suffix Arrays [23]. None of these pattern-matching approaches (DTW, ED, string kernels) require the size of the query pattern to be the same as the size of the sub-sequence.

A number of pattern-matching approaches require segmentation as a pre-processing step. These methods include the template-based (TB) [24], rule-based (RB) [24], hybrid (HY) [25] and decision tree (DT) approaches, which are all discussed in Section 2.4. All of these pattern matching approaches require the size of the query pattern to be the same as the size of the sub-sequence.

To reduce the number of data points in the original time series, segmentation methods have been commonly used as a pre-processing step in time series

* Corresponding author. Tel.: +853 88224454.
E-mail addresses: mb25466@umac.mo (Y. Wan), amoonfana@qq.com (X. Gong), fstasp@umac.mo (Y.-W. Si).

http://dx.doi.org/10.1016/j.asoc.2015.10.012
1568-4946/© 2015 Elsevier B.V. All rights reserved.

analyses. These segmentation methods include the perceptually important points (PIP) method [26], the piecewise aggregate approximation (PAA) method [27], the piecewise linear approximation (PLA) method [28] and the turning points (TP) method [29]. Three variations of the PIP method were compared by Fu et al. [24]. They considered the vertical distance PIP (PIP-VD) to be the best choice.

In using these segmentation methods, analysts must consider that each method involves a different process for the selection of data points from the input time series. Therefore, the resulting time series can be significantly different, depending on the segmentation method used. Such differences in segmented time series can have a profound effect on the pattern matching results. A number of studies have been conducted on the effects of segmentation on pattern matching methods. Chen et al. [30] compared the PIP-based evolutionary pattern discovery approach with the discrete wavelet transformation (DWT) approach (which is based on the pattern discovery approach). Their proposed approach solved the problems of information loss, distortion of segments and generation of meaningless patterns that had been associated with the DWT-based approach. Fu et al. [24] compared efficiency and effectiveness among several pattern matching approaches, namely the TB approach (which uses the PIP-VD segmentation method), the RB approach (which also uses the PIP-VD) and the PAA-based approach proposed by Keogh et al. [28]. The two approaches based on PIP performed better than the PAA-based approach. The TB approach provided an effective method, but the RB approach showed a better ability to describe query patterns. Zhang et al. [25] compared the processing time and accuracy of the proposed HY approach with that of two other pattern matching approaches the ED-based method and the slope-based method. The experimental results showed that the HY approach was more effective and efficient than the ED-based method or the slope-based approaches.

Comparisons of the PLA, PAA and PIP segmentation methods in terms of on-line use, representation interval and complexity were discussed by Si and Yin [29]. They reported that the PAA approach could be directly used for on-line representation, but the PIP and PLA approaches were unsuitable. PAA is based on segments with identical length calculations, but PIP and PLA are based on the degree of fluctuation in the time series. The complexity rating of PAA is $O(n)$, and the ratings of PIP and PLA are both $O(n^2)$. PAA is similar to the operation of removing redundant information from triangle meshes [31].

Si and Yin [29] also compared a segmentation method based on TPs with two common segmentation methods, PLA and PIP, in terms of capacity for reconstructing error and ability to keep trends. The PLA approach produced the least amount of errors and the fewest trends, and the proposed TP approach preserved more trends than the PLA or the PIP approaches. All of these studies, however, investigated only one or two pattern matching approaches each. To the best of our knowledge, no comparative analysis has been ever performed on all of the well-known methods of data segmentation methods and pattern matching.

In this paper, we evaluate the effectiveness and accuracy of four well-known approaches to pattern matching (the TB, RB, HB and the DT approaches) when used with four segmentation methods as the pre-processing step. These four segmentation methods are the PIP method [26], the PAA method [27], the PLA method [28] and the TP method [29]. We use the technical pattern known as H&S as a query pattern to better understand how these segmentation methods affect the pattern matching approaches.

The remainder of the paper is organised into four sections. We briefly review the algorithms used for segmentation and pattern matching in financial time series in Section 2. In Section 3, we report the experimental results obtained from evaluation of segmentation and pattern matching algorithms applied to price data from the Hong Kong stock market. In Section 4, we summarise our findings and discuss directions for future research.

2. Segmentation methods and pattern matching approaches

2.1. Terminology and notation

The term "time series" is defined as an ordered list $T = [(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n)]$. $Len(T)$ represents the number of points in T. As the value t_i is sequential (e.g., 1, 2,..., n), T can be simplified to $T = [x_1, x_2, \ldots, x_n]$ and T_i is often used to denote the element x_i. Accordingly, the sub-sequence S of T is $T_{i,j} = [x_i, x_{i+1}, \ldots, x_j]$, where i and j are the start and end points of the sub-sequence.

2.2. Segmentation methods

The aforementioned four pattern matching approaches (the TB, RB, HY and DT approaches) all require a pre-processing of the input sequence to reduce the number of data points until the length of the input sequence is the same as the query pattern. The well-known segmentation method, PIP, was first introduced by Chung et al. [26]. The variants of PIP are PIP-ED, PIP-VD and perpendicular distance PIP. Fu et al. [24] considered the PIP-VD method to be the best choice among these three variants in terms of efficiency and effectiveness. Therefore, we choose the PIP-VD method for our experiment. The generic algorithm of PIP is described in Algorithm 1 [24]. With the time series T, the first and the last data point in the time series are the first two PIPs. The third PIP is the point in T with maximum vertical distance to the line joining the first two PIPs. The fourth PIP is the point in T with maximum vertical distance to the line joining its two adjacent PIPs, either between the first and second PIPs or between the second and the last PIPs. This process continues until the length of the segmentation sequence SP is equal to the input sequence Q. An illustration showing the selection of five PIPs from a time series is shown in Fig. 1.

Algorithm 1. Pseudocode of the PIP identification [24]

```
  Function: PIP Identification (T, Q)
Input: sequence T of Len(T) = m, template Q of Len(Q) = n
Output: pattern SP of Len(SP) = n
  Set $SP_1 = P_1$, $SP_n = P_m$
  repeat
    Select point $T_j$ with maximum distance to the adjacent points in SP
  ($SP_1$ and $SP_n$ initially)
    Add $T_j$ to SP
  until all SP are all filled
  return SP
```

The PAA method was proposed by Keogh and Pazzani [27]. In PAA, a time series T of length n is represented by the compressed time series T' of length N. That is, $T = (x_1, \ldots, x_n)$ is represented by $T' = (x'_1, \ldots, x'_N)$. The time series T is divided into N equal-sized parts and each part is represented by the mean value of the data points in that part. The ith element of T' can be calculated by using Eq. (1).

$$x'_i = \frac{N}{n}\sum_{j=s_i}^{e_i} x_j \tag{1}$$

where s_i and e_i denote the start point and end point of the ith part, respectively. An illustration showing the selection of five points with PAA is shown in Fig. 2.

The PLA method uses several straight lines to segment a time series T. PLA can be obtained through the sliding window, top-down or bottom-up methods [28]. In our experiment, we choose the bottom-up method for obtaining PLA. The generic bottom-up algorithm [28] for PLA is described in Algorithm 2. In this bottom-up method, the time series is represented by a number of segments in the first FOR loop. The costs of merging the neighbour segments are calculated in the next FOR loop. In the WHILE loop, the two neighbouring segments with the lowest merge-costs are combined until the minimum merge cost is less than the threshold. In this experiment, the merging process continues until the number of data points in the sequence is equal those in the query pattern. Fig. 3 gives an illustration showing the selection of five points with PLA.

Algorithm 2. The generic algorithm of the PLA-bottom up [28]

```
Function: Seg_TS = Bottom_UP(T, max_error)
for i = 1: 2: Len(T) {Create initial fine approximation.} do
  Seg_TS = concat (Seg_TS, create_segment ($T_{i,i+1}$));
end for
for i = 1: Len(Seg_TS)-1 { Find the cost of merging each pair of segments.}
do
  merge_cost (i) = calculate_error ([merge (Seg_TS(i), Seg_TS(i +1))]);
end for
while min (merge_cost)< max_error {While not finished.} do
  i = min (merge_cost); { Find the cheapest pair to merge.}
  Seg_TS(i) = merge (Seg_TS (i), Seg_TS(i + 1))); {Merge them.}
  delete (Seg_TS(i + 1)); {Update records.}
  merge_cost (i) = calculate_error (merge (Seg_TS(i), Seg_TS(i + 1));
  merge_cost (i-1) = calculate_error (merge (Seg_TS(i-1), Seg_TS(i));
end while
```