Expert Systems With Applications 55 (2016) 284–296

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Financial time series pattern matching with extended UCR Suite and Support Vector Machine

Xueyuan Gong, Yain-Whar Si*, Simon Fong, Robert P. Biuk-Aghai

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

ARTICLE INFO

Keywords:
Financial time series
Subsequence matching
Perceptually important points
UCR Suite
Support Vector Machine

ABSTRACT

Chart patterns are frequently used by financial analysts for predicting price trends in stock markets. Identifying chart patterns from historical price data can be regarded as a subsequence pattern-matching problem in financial time series data mining. A two-phase method is commonly used for subsequence pattern-matching, which includes segmentation of the time series and similarity calculation between subsequences and the template patterns. In this paper, we propose a novel approach for locating chart patterns in financial time series. In this approach, we extend the subsequence search algorithm UCR Suite with a Support Vector Machine (SVM) to train a classifier for chart pattern-matching. The experimental results show that our approach has achieved significant improvement over other methods in terms of speed and accuracy.

1. Introduction

A time series is a sequence of data points that are typically measured at successive, uniform time intervals. Time series are used in numerous areas such as the representation of the daily closing price of stocks, electrocardiograms, the gait of a walking person, and speech patterns. In general, time series pattern matching can be categorised into whole-sequence matching and subsequence matching (Agrawal, Faloutsos, & Arun, 1993). Whole-sequence matching focuses on the similarities between the time series (sequence) and the pattern, whereas subsequence matching addresses issues of similarity between the pattern and a subsequence within the sequence. Note that a subsequence of a sequence can be treated as a shorter sequence, and likewise, a pattern can also be considered as a kind of sequence that is associated with specific characteristics or meaning. Thus, the similarity calculation between a subsequence and a pattern can be formulated as the similarity calculation between two sequences.

Subsequence pattern matching has been applied to numerous areas, such as the location of technical patterns in financial time series in deciding when to buy or sell stocks, or the search for anomalies in time series in the health care domain for the detection of diseases. In the area of financial time series, patterns are also known as chart patterns and are widely used for technical analysis of the stock market. Specifically, a chart pattern is formed within a chart when prices are graphed. Some of the common chart patterns (Lo, Mamaysky, & Wang, 2000) that are widely used by traders and investors are depicted in Fig. 1.

Due to their high value in financial analysis, these chart patterns are commonly adopted as test patterns in many pattern-matching methods (Fu, Chung, Luk, & Ng, 2007; Zhang et al., 2010). Therefore, subsequence pattern matching in financial time series can be considered as the task of locating the subsequence of a time series with a shape similar to that of a query pattern. An example of a subsequence and the corresponding matched pattern is depicted in Fig. 2.

Many approaches in the financial time series area make use of perceptually important points (PIPs) (Chung, Fu, Luk, & Ng, 2001) to perform segmentation on subsequences as a pre-processing step, followed by pattern matching on the segmented subsequences with template-based (TB) (Fu et al., 2007), rule-based (RB) (Fu et al., 2007), and Hybrid (Chung et al., 2001) approaches. However, segmentation (with PIP or other methods) has three disadvantages:

- The quality of segmentation is important. The important features of the original subsequence should be retained and insignificant noise should be eliminated. If segmentation methods cannot ensure a good outcome, the pattern-matching step that follows can produce a poor result. An example of the selection of the wrong features by the segmentation method in the generation of a subsequence is depicted in Fig. 3. X in Fig. 3(a) represents the segmented subsequence of S by PIP; we can see that X is not at all similar to the pattern P. However, S should

* Corresponding author. Tel.: +853 8822 4454; fax: +853 8822 2426.
E-mail addresses: amoonfana@qq.com (X. Gong), fstasp@umac.mo (Y.-W. Si), ccfong@umac.mo (S. Fong), robertb@umac.mo (R.P. Biuk-Aghai).

http://dx.doi.org/10.1016/j.eswa.2016.02.017
0957-4174/

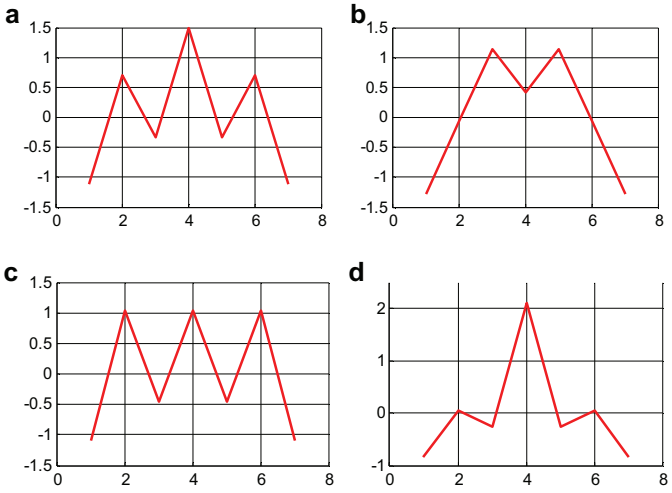

Fig. 1. Common chart patterns (a) Head&Shoulders, (b) DoubleTop, (c) TripleTop, and (d) SpikeTop.

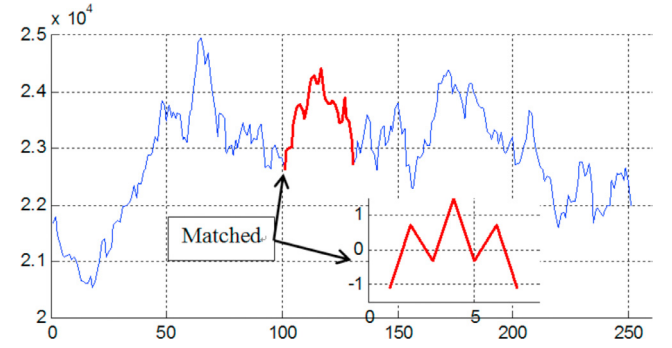

Fig. 2. Subsequence pattern matching in financial time series.

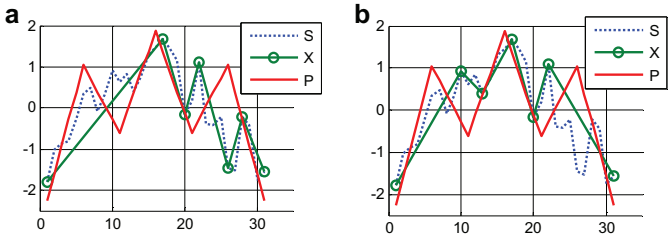

Fig. 3. Example for illustration of the possible effects of segmentation on the outcome of pattern matching, (a) result of PIP and (b) expected result of PIP.

be similar to P, because if S is segmented like X in Fig. 3(b), then S is similar to P. This shows that the performance of segmentation affects the accuracy of subsequence matching.

- Information loss can affect the pattern-matching result. Segmentation means sacrificing accuracy for reduced computation time. However, the use of an insufficient number of points can cause information loss and indivertibly affect the pattern-matching result. For instance, in previous financial time series pattern-matching analyses (Fu et al., 2007; Zhang et al., 2010), the patterns have been represented by only seven points. For a subsequence of length 31 and a pattern of length 7, the compression rate reaches $(31-7)/31 = 0.774$, which means that 77.4% of the original subsequence points are eliminated by segmentation. Therefore, the information retained by the remaining 22.6% of the points is fewer than the original subsequence.
- Time differences must be taken into account after the segmentation. Without segmentation, the subsequence and pattern are point-to-point, namely, the time values (x) of the corresponding price values (y) between the subsequence and the pattern are the same.

All of these discussions illustrate that segmentation as a pre-processing step can cause serious false-positive and false-negative results in the pattern-matching process. In this paper, we first compare the subsequence pattern-matching approaches from two categories, those that require segmentation as a pre-processing step and those that do not. Based on the results of the comparison, we propose a novel approach called the extended UCR Suite (EUCRS) that comprises two components: the UCR Suite (UCRS) and a Support Vector Machine (SVM). The EUCRS is an extended version of the UCRS (Keogh, 2002) developed by Rakthanmanon.

1.1. Definitions and notations

All definitions and notations used in this paper are defined in this section. For the purpose of illustration, we treat each time series as an ordered set. $T = \{(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n)\}$ represents a time series T that starts at time t_1 and ends at t_n. The length of T is $|T|$, namely, the number of elements n in T, written as $|T| = n$. Each element T_i of T, denoted as (t_i, x_i), is the value x_i at time t_i. Because the time value t_i in financial time series is sequential and often ignored, we can simplify the time series into $T = \{x_1, x_2, \ldots, x_n\}$. Accordingly, t_i can be simply replaced by the value of subscript i, namely $t_1 = 1$, $t_2 = 2$, ..., $t_n = n$. X is the segmentation of T. It usually contains fewer elements than T, and the segmentation is usually performed within the range of acceptable information loss.

A subsequence of T starting at t_i and ending at t_j is written as $T_{i,j} = \{(t_i, x_i), (t_{i+1}, x_{i+1}), \ldots, (t_j, x_j)\}$, or it can be simplified to $T_{i,j} = \{x_i, x_{i+1}, \ldots, x_j\}$, where $1 \le i < j \le n$. Note that like $T_{i,j}$, X_{ij} is also a subsequence of T, but $X_{i,j}$ is discrete on most occasions; we will introduce that in detail in Section 3.2.1. In addition, the pattern P is also a time series.

1.2. Problems of pattern matching

The problem of financial time series pattern matching can be divided into five steps,

1. *Define patterns*: to the best of the authors' knowledge, none of the technical patterns have an exact definition. Line charts are mostly used to represent technical patterns. Therefore, the criteria by which we should define a pattern and store it in a computer are ambiguous. For a TB approach, the pattern is defined as a time series. For an RB approach, the pattern is defined as a set of rules. We will discuss these approaches in more detail in Section 3.1.1.
2. *Subsequence search*: as mentioned in Fu et al. (2007), there is no need to find all subsequences $T_{i,j}$, because analysts and traders only care about a pattern with proper length, e.g., $28 < |T_{i,j}| < 32$. For the traders, monitoring the appearance of the latest $T_{i,j}$ that is similar to pattern P is more important than finding all of the subsequences. The aim of the subsequence search step is to find a $T_{i,n} = \{x_i, x_{i+1}, \ldots, x_n\}$ that is similar to pattern P, where $|T_{i,n}| = m$, and m is a user-specified size. The procedure that details the subsequence search is introduced in Section 3.1.2.
3. *Segmentation*: the volume of the time series data is too large for current methods to calculate; this is known as the "dimensionality curse" (Fu, Chung, Luk, & Ng, 2008). In addition, because a comparison of two time series with different lengths is not convenient, segmentation is commonly used for calculation of