Engineering Applications of Artificial Intelligence 26 (2013) 2581–2596

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

OBST-based segmentation approach to financial time series

Yain-Whar Si*, Jiangling Yin

Department of Computer and Information Science, University of Macau, Av. Padre Tomas Pereira, Taipa, Macau

ARTICLE INFO

Article history:
Received 2 February 2013
Received in revised form
26 June 2013
Accepted 29 August 2013
Available online 7 October 2013

Keywords:
Financial time series
Turning points
Segmentation
Trends
Optimal binary search tree

ABSTRACT

Financial time series data are large in size and dynamic and non-linear in nature. Segmentation is often performed as a pre-processing step for locating technical patterns in financial time series. In this paper, we propose a segmentation method based on Turning Points (TPs). The proposed method selects TPs from the financial time series in question based on their degree of importance. A TP's degree of importance is calculated on the basis of its contribution to the preservation of the trends and shape of the time series. Algorithms are also devised to store the selected TPs in an Optimal Binary Search Tree (OBST) and to reconstruct the reduced sample time series. Comparison with existing approaches show that the time series reconstructed by the proposed method is able to maintain the shape of the original time series very well and preserve more trends. Our approach also ensures that the average retrieval cost is kept at a minimum.

1. Introduction

Time series is a temporal data object which is used to store a collection of observations made chronology (Fu, 2011). Time series are usually generated from a wide range of scientific and financial applications. Examples of time series include historical price and trading volume obtained from financial stock markets, motion and speech data from computer vision applications, and electrocardiogram (ECG) data from medical diagnosis systems. Time series are often characterised as large in data size, high in dimensionality, and require continuous updating (Fu, 2011).

There are two common approaches for analysing financial markets; fundamental analysis and technical analysis. In fundamental analysis, factors such as profit and loss, performance data, and analysis for the corresponding industry are used to predict the movement of the price. Analysts may also take into account other factors such as government policies and political climate for the prediction. However, fundamental analysis often fails to predict price movement due to the sheer number of factors involved and the influence of the biases from the analysts' judgment. In technical analysis, time series data such as historical price, volume, and other statistical data are used to predict the future price. In this paper, we address the segmentation problem of financial time series.

Segmentation is often performed prior to locating technical patterns such as Head-and-Shoulder or Double Tops patterns in financial time series. In order to match with the patterns from the pattern template in different resolution, feature points need to be extracted during the segmentation process and matched with the pattern template (Fu et al., 2007; Zhang et al., 2010a). For instance, an extended version of Perceptually Important Points (PIP) algorithm is used to segment the time series for locating technical patterns such as Up-Trapeziform, Bottom-Trapeziform, Up-flag, and Down-flag from the bid and trade time series of Chicago stock market (Zhang et al., 2007). Segmentation method such a Piecewise Linear Regression method is also used to find the past trends in the stock data (Lavrenko et al., 2000). The trends identified in the segmentation step are then used in combination with the news stories for predicting the future trends. Piecewise Linear Representation method is also used as a pre-processing step to segment the time series before predicting the future stock price based on a multilayer feed forward artificial neural network (Kwon and Sun, 2011).

Various methods for the time series segmentation have been proposed, including Discrete Fourier Transform (DFT) (Agrawal et al., 1993; Chu and Wong, 1999), Discrete Wavelet Transform (DWT) (Chan and Fu, 1999; Kahveci and Singh, 2001), Piecewise Aggregate Approximation (PAA) (Keogh et al., 2001a), Adaptive Piecewise Constant Approximation (APCA) (Chakrabarti et al., 2002), Piecewise Linear Representation (Piecewise Constant Approximation) (Keogh, 1997; Keogh and Pazzani, 2000), and Singular Value Decomposition (SVD) (Kanth et al., 1998). These segmentation approaches usually focus on the lower bound of the Euclidean distance, but smooth out the salient points of the original time series (Fu et al., 2008). Such smoothing effect inadvertently removes the important points which are crucial in identifying the change in trends and shapes. In this paper, we propose an approach for identifying and storing important points

* Corresponding author. Tel.: +853 83974923.
E-mail addresses: fstasp@umac.mo (Y.-W. Si), jyin@eecs.ucf.edu (J. Yin).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.08.015

which can be used to preserve the overall shape and trend of the time series.

Uncovering financial market cycles and forecasting market based on trends has been extensively studied by analysts. One of the renowned studies on trend analysis is the Elliott Wave Principle (Frost and Prechter, 2001) which is used to describe movement of the stock based on five distinct waves on the upside and three distinct waves on the downside. The major and minor waves determine the major and minor market trends, respectively. Elliott Wave Principle becomes one of the important foundations in stock market analysis for identifying significant trends for forecasting price movement. When analysing or predicting the movement of financial time series, analysts often rely on the trends contained within the time series rather than on the exact value of each data point in that series (Atsalakis and Valavanis, 2009; Chung et al., 2002; Kamijo and Tanigawa, 1990). In addition, Dow Theory (Hamilton, 2010) also categorises stock movement into primary and secondary trends. Accordingly, as an important characteristic, these trends should be taken into consideration during time series segmentation. In this paper, we propose a segmentation method which is not only capable of segmenting the time series for storing and incremental retrieving of important data points but also capable of preserving as much trends as possible for further data mining tasks.

A time series comprises numerous data points. However, the fluctuations, trends and patterns within a time series are reflected by only a small number of local maxima or minima data points. These data points in time series are often called landmarks and can be defined by the points, times or events of greatest importance, with the points primarily part of the local maxima or minima. Perng et al. (2000) define a landmark model for performing similarity pattern searching in a database. Instead of using raw data, the model proposed in Perng et al. (2000) employs landmarks to improve the efficiency of series data processing. The perceptually important points (PIP) method proposed in Zhang et al. (2010a) also selects critical points called PIPs to preserve the general shape of the time series. The resulting PIPs are able to reflect any fluctuation from the original time series. Keogh et al. (2001b) use the end points of the fitted line segments to compress the time series, whereas Pratt (2001) also select part of the local maxima or minima to perform a similar search for time series. In this paper, we propose the use of the local minimum and maximum points on a time series, called Turning Points (TPs), for segmentation. TPs can be extremely useful in technical analysis because they can be used to preserve the overall shape and trend of the time series compared to other data points.

To allow the incremental reconstruction of reduced-sample time series from these TPs at a later stage, we propose an algorithm to calculate the degree of importance for each TP, which is a measure used to compare its contribution to the preservation of the trends and overall shape of the original time series. Using the degree of importance as a measure allows TPs to be ordered on the basis of their priority. An Optimal Binary Search Tree (OBST) is then used to store the TPs. The advantage of the proposed storage scheme is that it allows the more important points to be retrieved at an early stage to reconstruct the reduced-dimension time series. The use of an OBST also guarantees that the average search cost is kept to a minimum.

In addition, analysts can employ the proposed method to reconstruct reduced-dimension time series at different detail levels by retrieving TPs based on their degree of importance. Such a property allows the top-down analysis of the time series in which highly visible trends are first identified and allows the retrieval of more detailed segments to be postponed until later stages. Our experimental findings show the proposed method to achieve promising results in preserving a higher number of trends than existing methods.

The paper is an extension of our previous conference paper (Yin et al., 2011) and as such the background and definition of turning point are taken from this previous work. The remainder of the paper is organised into five sections. We discuss related work on financial time series segmentation in Section 2. The algorithms for calculating the degree of importance and storing the TPs into an OBST are provided in Section 3. In Section 4, we report details of the experimental results obtained from tests of our proposed approach with price data from the Hong Kong stock market, and in Section 5, we summarise our ideas and discuss directions for future research.

2. Related work

The multitude of time series data produced by a wide range of applications has been increased at an unprecedented pace in recent years. Application such as financial trading, medical diagnosis, computer vision, and speech recognition can produce a vast amount of time series data. The complexity and dimensionality of the time series data vary from one application to another. For instance, time series data from computer vision applications such as body pose tracking and action recognition are in high dimension and non-linear.

A number of techniques have been proposed by the scientific community to analyse these high dimensional time series data. Dimensionality reduction is often referred to as a feature extraction step in machine learning and statics. For instance, linear dimensionality reduction method such as Principal Component Analysis (PCA) (Pearson, 1901) transforms the data from high-dimensional space to low-dimensional space. PCA has been widely used in applications such as face recognition to reduce the number of variables.

Two main categories of techniques are available in non-linear dimensionality reduction methods; mapping-based and embedding-based methods (Lewandowski et al., 2010). Mapping-based approach such as Gaussian Process Latent Variable Mode (GP-LVM) (Lawrence, 2003), Back Constraint Gaussian Process Latent Variable Mode (BC-GPL-VM) (Lawrence and Quinonero-Candela, 2006), Gaussian Process Dynamical Model (GPDM) (Wang et al., 2006) comprise a mapping from a latent space to the data space. GP-LVM technique proposed by Lawrence (2003) is a non-linear generalisation of probabilistic PCA providing a smooth mapping from latent to data space. Embedded-based approaches such as Laplacian Eigenmaps (LE) (Belkin and Niyogi, 2001), Temporal Laplacian Eigenmaps (TLE) (Lewandowski et al., 2010), Isomap (Tenenbaum et al., 2000), and ST_Isomap (Jenkins and Mataric, 2004) estimate the structure of the underlying manifold by approximating each data point according to their local neighbours on the manifold (Lewandowski et al., 2010).

Financial time series analysis and prediction constitute active research problems in the data mining area due to the attractive commercial applications and benefits that data mining offer (Kamijo and Tanigawa, 1990; Ralanamahatana et al., 2005). A number of algorithms, such as classification, clustering, segmentation, indexing and rule discovery (Atsalakis and Valavanis, 2009; Chung et al., 2002) from historical time series have been introduced in this area.

The segmentation of financial time series is a crucial preprocessing step in financial time series analysis. A number of approaches have been proposed for sampling and segmentation of financial time series: Discrete Fourier Transform (DFT) (Agrawal et al., 1993; Chu and Wong, 1999), Discrete Wavelet Transform (DWT) (Chan and Fu, 1999; Kahveci and Singh, 2001), the Piecewise Linear, and Piecewise Aggregate Approximation (PAA) (Keogh et al., 2001a, 2001b), and Singular Value Decomposition (SVD)