Expert Systems with Applications 40 (2013) 1579–1590

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A SAX-GA approach to evolve investment strategies on financial markets based on pattern discovery techniques

António Canelas, Rui Neves, Nuno Horta *

Instituto de Telecomunicações, Instituto Superior Técnico, Torre Norte, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal

ARTICLE INFO

Keywords:
Pattern discovery
Frequent patterns
Pattern recognition
Financial markets
Time series
Genetic algorithm
SAX representation

ABSTRACT

This paper presents a new computational finance approach, combining a Symbolic Aggregate approXimation (SAX) technique together with an optimization kernel based on genetic algorithms (GA). The SAX representation is used to describe the financial time series, so that, relevant patterns can be efficiently identified. The evolutionary optimization kernel is here used to identify the most relevant patterns and generate investment rules. The proposed approach was tested using real data from S&P500. The achieved results show that the proposed approach outperforms both B&H and other state-of-the-art solutions.

1. Introduction

The domain of computational finance has received an increasing attention by people from both finance and intelligent computation domains. The main driving force in the field of computational finance, with application to financial markets, is to define highly profitable and less risky trading strategies. In order to accomplish this main objective, the defined strategies must process large amounts of data which include financial markets time series, fundamental analysis data, technical analysis data, etc. and produce appropriate buy and sell signals for the selected financial market securities. What may appear, at a first glance, as an easy problem is, in fact, a huge and highly complex optimization problem, which cannot be solved analytically. Therefore, this makes the soft computing and in general the intelligent computation domains specially appropriate for addressing the problem. Recently, several works like (Krause, 2011; Parque, Mabu, & Hirasawa, 2010; Parracho, Neves, & Horta, 2011; Pinto, Neves, & Horta, 2011), have been published in the field of computational finance where soft computing methods are used for stock market forecasting, however, due to the complexity of the problem and the lack of generalized solutions this is undoubtedly an open research domain.

The use of chart patterns is widely spread among traders as an additional tool for decision making, however, the problem in this case is to say how close enough should the market match a specified chart pattern to make a buy or sell decision. In this paper a new approach combining a Symbolic Aggregate approXimation (SAX) technique together with an optimization kernel based on genetic algorithms (GA) is presented. The SAX representation is used to describe the financial time series, so that, relevant patterns can be efficiently identified. The evolutionary optimization kernel is here used to identify the most relevant patterns and generate investment rules. The proposed approach was tested using real data from S&P500. Finally, the achieved results outperform the existing state-of-the-art solutions.

This paper is organized as follows; in Section 2 the related work is discussed. Section 3 describes the method of dimensional reduction of the time series used in the paper, SAX. Section 4 the proposed approach that puts together the GA and SAX is explained. Section 5 describes the experiments and discusses results. Section 6 draws the conclusions.

2. Related work

First of all a distinction between pattern recognition and pattern discovery should be made. Recognition is identifying some patterns that are known on the time series, this case is a supervised approach, where a library of patterns (Bulkowski, 2005), is created and is made a search on the data market, trying to identify them (Parracho et al., 2011). In pattern discovery the quest is to find new patterns that occur in the time series, in this case, typically some data segments or windows are compared with others. This case is an unsupervised approach, which is also the case presented on this paper.

Prediction of financial markets has been subject of many studies. In this last few years a combination of algorithms and methods have been used, Table 1. Many of the applications use GA, proving the good results of this type of optimization tool in the financial market world.

* Corresponding author.
E-mail addresses: antonio.canelas@ist.utl.pt (A. Canelas), rui.neves@ist.utl.pt (R. Neves), nuno.horta@lx.it.pt (N. Horta).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.09.002

Table 1
Investment algorithms.

Ref.	Year	Method	Used data	Financial market	Period	Algorithm performance
Matsui and Sato	2010	GA	Several	Nikkei 225	Jan. 1999–Dec. 2009	57.4% (profit rate)
Hao	2010	GA-ANN	Stock price	Shenzhen	N/A	0.7176 (correlation between prediction and actual value)
Ng, Liang, Chan, and Yeung	2011	MCS RBFNN	Stock price	Hang Seng Index	10 years	167 (average earning)
Tahersima, Tahersima, Fesharaki, and Hamedi	2011	NN	Price	FOREX	September 20, 2010 to January 21, 2011	0.666186e-3-0.101144e-2 (mean square error)
Wang and Wang	2011	HLP-ANN	Stock price	Shanghai index	11-18-1991 to 2-10-2009	7.16–9.65% (prediction error)
Fang, Luo, Fei, and Li	2010	Wavelet Modulus Maxima + Kalman Filter	Daily trading volume	Bowin Technology & Denghai seed industry	February 13, 2008 to February 13, 2009	0.13–0.1879 (SNR – prediction)

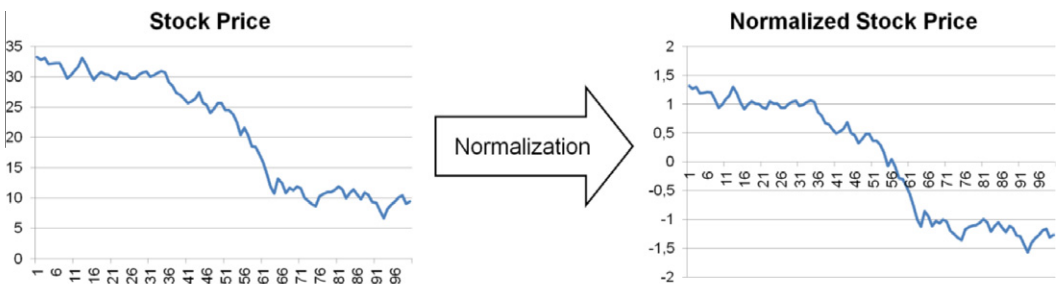

Fig. 1. Normalization process of the stock quote time series.

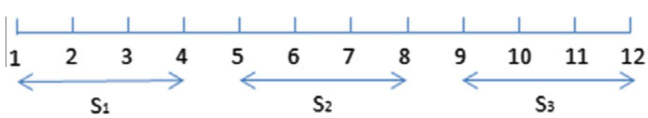

Fig. 2. Size 12 window divided in 3 segments, each point contributes to one segment only.

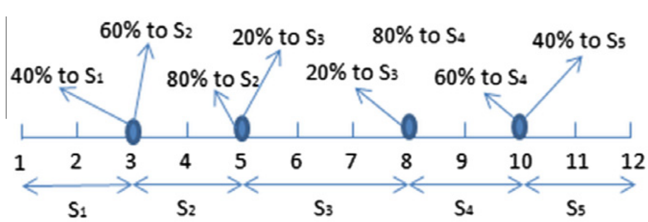

Fig. 3. Size 12 window divided in five segments, the points between segments contribute to the neighbors segments.

In order to create an efficient method of search, the time series should suffer some dimensionality reduction, the method used in this transformation must preserve the key essence of the data. Some of the methods to achieve this goal are the more commonly Discrete Fourier Transform (DFT) (Agrawal, Faloutsos, & Swami, 1993), Perceptually Important Points (PIP) (Fu, Chung, Kwok, & Ng, 2008), Piecewise Aggregate Approximation (PAA) (Keogh, Chakrabarti, Pazzani, & Mehrotra, 2001).

More recently, methods of symbolic representation of data and dimensional reduction began to appear, one of this methods is Symbolic Aggregate approXimation (SAX) (Lin, Keogh, Lonardi, & Chiu, 2003), which is based on PAA. This algorithm begins to divide the time series in windows, then each window in segments and reduces a set of points in each segment to their arithmetic mean and then converts this value to a symbol. To search patterns the sequences of symbols must be compared with each other to find similarities in the data, in the next section this method will be described in detail.

3. SAX method

In order to find patterns, large time series of dimension m will be break into smaller time series windows of size $n \ll m$. These windows must be compared with each other, so the characteristics of these time series must be similar, same magnitudes and base line. Therefore to apply this transformation to the windows, data has to be normalized (1), this normalization does not affect the original shape (Goldin & Kanellakis, 1995) and scales the data to the same relative magnitude Fig. 1.

$$x'_i = \frac{x_i - \mu_x}{\sigma_x} \tag{1}$$

where, x_i is the Points in window W_k; μ_x is the Mean of the points in W_k; and σ_x is the Standard deviation of all the x_i.

After normalization the data windows are ready to be compared, but the dimension of this data is high. At this point no data has been removed from the original time series, turning this process very expensive in time and computational resources. So some method of dimensionality reduction is needed, as said before SAX is based on PAA to achieve this objective.

In PAA the time series windows are divided in w equal size segments and each segment is represented by the arithmetic mean of the points in it, according to (2).

$$\bar{x}_j = \frac{w}{n} \sum_{i=\frac{n}{w}(j-1)+1}^{\frac{n}{w}j} x'_i \tag{2}$$

where, w is the number of segments; n is the size of the window and x'_i is the data point in the window.

This Eq. (2) is valid if n/w has an integer result, in this case each point contribute entirely to the frame where is inserted, Fig. 2.

In the case of a non integer relation, the point in the frontier between segments must contribute with some part to each of the