Physica A 416 (2014) 183–191

Contents lists available at ScienceDirect

Physica A

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

Financial time series analysis based on information categorization method

Qiang Tian *, Pengjian Shang, Guochen Feng

Department of Mathematics, School of Science, Beijing Jiaotong University, No. 3 of Shangyuan Residence, Haidian District, Beijing 100044, People's Republic of China

HIGHLIGHTS

- We apply the information categorization method and succeed to classify the financial markets.
- We investigate the similarity among US and Chinese stock indices for the Asian currency crisis and the global financial crisis.
- The phylogenetic trees of different stock indices are presented and provide obvious phenomenon.

ARTICLE INFO

Article history:
Received 29 May 2014
Received in revised form 26 July 2014
Available online 30 August 2014

Keywords:
Information categorization method
Rank order
Dissimilarity index
Financial time series

ABSTRACT

The paper mainly applies the information categorization method to analyze the financial time series. The method is used to examine the similarity of different sequences by calculating the distances between them. We apply this method to quantify the similarity of different stock markets. And we report the results of similarity in US and Chinese stock markets in periods 1991–1998 (before the Asian currency crisis), 1999–2006 (after the Asian currency crisis and before the global financial crisis), and 2007–2013 (during and after global financial crisis) by using this method. The results show the difference of similarity between different stock markets in different time periods and the similarity of the two stock markets become larger after these two crises. Also we acquire the results of similarity of 10 stock indices in three areas; it means the method can distinguish different areas' markets from the phylogenetic trees. The results show that we can get satisfactory information from financial markets by this method. The information categorization method can not only be used in physiologic time series, but also in financial time series.

1. Introduction

Recently, financial markets have become active areas and attracted much attention; they are remarkably well-defined complex systems with a large number of interacting units that conform to the underlying economic trends. Physicists are currently contributing to the modeling of complex systems by using tools and methodologies developed in statistical mechanics and theoretical physics. Econophysics [1–3] is the term used to denote the application of statistical mechanics to economic systems. A range of methods have been introduced to investigate stock markets as a reflection of economic trends. The similarity between financial time series is an important feature of the dynamics of financial markets.

Many different methods have been used to quantify the similarity in stock time series. After Pincus introduced the approximate entropy (ApEn) to quantify the concept of changing complexity [4–6], some authors used the ApEn to measure

* Corresponding author. Tel.: +86 15210584752.
E-mail address: 13121599@bjtu.edu.cn (Q. Tian).

http://dx.doi.org/10.1016/j.physa.2014.08.055
0378-4371/

the biologic time series [7,8]. Moreover, Richman et al. analyzed the shortcomings of the ApEn method and developed sample entropy (SampEn) over a broad range of conditions, it was widely applied in clinical cardiovascular studies [9,10]. Then Cross-ApEn and Cross-SampEn were introduced to measure the similarity of two distinct time series [11]. These methods were constructed based on quantifying the regularity of time series, and initially aimed at estimating the system complexity of stock markets. Also, Costa et al. introduced the multiscale entropy to take into account the time scales and applied it to measure the complexity of biologic systems [12–14]. Based on multiscale entropy and the Cross-SampEn, multiscale cross-sample entropy was proposed to analyze the similarity of two series under different time scales [15]. For further analysis, multiscale time irreversibility was proposed to classify the financial markets and got nearly the same results [16]. Besides, there are many other methods to analyze the similarity of different stock markets, such as multiscale detrended fluctuation analysis (MSDFA) and multiscale detrended cross-correlation analysis (MSDCCA) [17], three-phase clustering method [18], and Time-Varying Copula–GARCH Model [19].

The measurement of the similarity [20,21] between two complex sequences was proposed by Yang et al. in 2003. They developed a novel information-based similarity index to detect and quantify hidden dynamical structures in the human heart rate time series using tools from statistical linguistics. They proposed a method in which the time series can be mapped to binary symbolic sequences and the dissimilarity index can be calculated through rank-frequency. They applied the method of constructing phylogenetic trees [22] to arrange different groups of samples on a branching tree to best fit the pair-wise distance measurements. For better consequences, Peng et al. proposed another definition for the weighting factor by using Shannon entropy [23], and they gave another definition of the dissimilarity index, then they applied the novel definition for an information categorization approach [24], biologic signals [25], SARS Coronavirus [26], and the patterns of blood pressure signals [27]. Besides, Peng et al. had proved that the modification of the dissimilarity index can provide the best classification across all types of symbolic sequences [25].

In previous studies, a number of works focused on the analysis of physiologic signals by using the above method. The measurement of similarity is the parameter of linguistic analysis algorithm and it is a kind of pattern analysis method. Based on this method, we take account of the situation of stock time series and choose the proper size of the word which is different from Peng et al. [25], and then we analyze the similarity of stock markets. However, most of previous methods are applied to the time series directly and the length of the time series should be same. The information categorization method is based on symbolic sequences, so we do not need to consider the length of sequences, just choose the proper m-bit words. The financial market dynamics are driven by a number of complex factors: index at the same level, the sub-index, the economic data, trading sentiment, trading price, weight, and other stock information. For this type of intrinsically noisy system, it may be useful to simplify the dynamics via mapping the output to binary sequences, where the increase and decrease in every day of the stock market closing prices are denoted by 1 and 0, respectively [28]. The resulting binary sequence retains important features of the dynamics generated by the underlying control system, but is tractable enough to be analyzed as a symbolic sequence. So we succeed in turning the complex stock time series to symbolic sequences, which is the first significant step to apply this method. The novelty of this information categorization method is that it incorporates elements of both information-based and word statistics-based categories since the rank order difference of each word statistics is weighted by its information content using Shannon entropy. Furthermore, the composition of these basic elements captures global information related to usage of respective elements in stock time series. Also, the phylogenetic trees based on the dissimilarity index can give us direct information about different markets to analyze the similarity among them.

The reminder of the paper is organized as follows. In the following section we present the details of information categorization method. In Section 3, we propose the method to analyze the similarity of the Chinese and US stock markets in different time periods and extend to three stock markets. Finally, we summarize the findings of this paper in the last section.

2. Method

The information categorization method was proposed by Peng et al. [25], it is the modification based on the method proposed by Yang et al. [20,21]. Based on this method, we consider the stock time series and choose the appropriate way to analyze the stock markets. Now we will briefly review the modified method.

Consider a financial market time series, $\{x_1, x_2, \ldots, x_N\}$, where x_i is the closing price in day i. We can classify each pair of successive closing prices into one of the two states that represents a decrease in x, or an increase in x. These two states are mapped to the symbols 0 and 1, respectively

$$I_n = \begin{cases} 0 & \text{if } x_n \le x_{n-1} \\ 1 & \text{if } x_n > x_{n-1}. \end{cases} \tag{1}$$

To define the measurement of similarity between two symbolic sequences, we carry out the following procedures. First, we map $m+1$ successive interval to a binary sequence of length m, called an m-bit word. Each m-bit word, w_k, therefore, represents a unique pattern of fluctuations in a given time series. By shifting one data point at a time, the algorithm produces a collection of m-bit words over the whole time series. Therefore, it is plausible that the occurrence of these m-bit words reflects the underlying dynamics of the original time series. Different types of dynamics thus produce different distributions of these m-bit words. Then we count the occurrences of different words, and then sort them in descending order by frequency of occurrence. The most frequently occurring word is ranked number 1, and so on.