Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
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.

© 2014 Elsevier B.V. All rights reserved.

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
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 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

\[
I_n = \begin{cases} 0 & \text{if } x_n \leq x_{n-1} \\ 1 & \text{if } x_n > x_{n-1}. \end{cases}
\]

(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.
Fig. 1. Rank order comparison of two stock time series from the same time period. For each word, its rank in the first time series is plotted against its rank in the second time series, the dashed diagonal line indicates the case where the rank order of words for both time series. Note the greater scattering of data points (dissimilarity) between time series 2 and 1 is compared to the time series 3.

The resulting rank-frequency distribution, therefore, represents the statistical hierarchy of symbolic words of the original time series. For example, the first rank word corresponds to one type of fluctuation which is the most frequent pattern in the time series. In contrast, the last rank word defines the most unlikely pattern in the time series. Note that for any m-bit word, its rank order can be different in these two sequences. Therefore, we can plot the rank number of each m-bit word in the first time series against that of the second time series (see Fig. 1).

If two time series are similar in their rank order of the words, the scattered points will be located near the diagonal line. Therefore, the average deviation of these scattered points away from the diagonal line is a measure of the distance between these two time series. Greater distance indicates less similarity and vice versa. In addition, we incorporate the likelihood of each word in the following definition of a weighted distance, $D_m$, between two symbolic sequences, $S_1$ and $S_2$.

$$D_m(S_1, S_2) = \frac{1}{2^m - 1} \sum_{k=1}^{2^m} |R_1(w_k) - R_2(w_k)| F(w_k)$$

where

$$F(w_k) = \frac{1}{Z} [-p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k)].$$

Here $p_1(w_k)$ and $R_1(w_k)$ represent probability and rank of a specific word, $w_k$, in time series $S_1$. Similarly, $p_2(w_k)$ and $R_2(w_k)$ stand for probability and rank of the same m-bit word in time series $S_2$. The absolute difference of ranks is multiplied by the normalized probabilities as a weighted sum by using Shannon entropy [4] as the weighting factor. Finally, the sum is divided by the value $2^m - 1$ to keep the value in the same range of $[0, 1]$. The normalization factor $Z$ is given by

$$Z = \sum_k [-p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k)].$$

3. Results and discussion

3.1. Analysis of financial time series in different time periods

We employ the information categorization method mentioned in Section 2 to classify complex signals. The analyzed dataset consists of six indices: three US stock indices, DJIA, NASDAQ, and S&P500 together with three Chinese stock indices, ShangZheng, ShenCheng and HSI. The Data A are recorded every day of closing prices from April 3, 1991, to December 31, 2013. And then we split the 6 daily stock indices into three periods, the first is from 1991 to 1998 (Data B), the second is from 1999 to 2006 (Data C) and the third is from 2007 to 2013 (Data D). Data B are obtained before the Asian currency crisis, Data C are obtained after the Asian currency crisis and before the global financial crisis, and Data D are obtained during and after the global financial crisis. The data were collected from the Yahoo Financial web site [29].
Table 1
Distances between all these time series from the year 1991 to the year 2013.

| Distance | ShangZheng | ShenCheng | HSI | S&P500 | NASDAQ | DJIA |
|----------|------------|-----------|-----|--------|--------|------|
| ShangZheng | 0.1905 | 0.2616 | 0.3011 | 0.2568 | 0.2561 |
| ShenCheng | 0.1905 | 0.2783 | 0.3443 | 0.2964 | 0.3532 |
| HSI | 0.2616 | 0.2783 | 0.2174 | 0.2217 | 0.2597 |
| S&P500 | 0.3011 | 0.3443 | 0.1775 | 0.1775 | 0.1769 |
| NASDAQ | 0.2568 | 0.2964 | 0.2217 | 0.1775 | 0.2435 |
| DJIA | 0.2561 | 0.3532 | 0.2597 | 0.1769 | 0.2435 |

Table 2
Distances between all these time series from the year 1991 to the year 1998.

| Distance | ShangZheng | ShenCheng | HSI | S&P500 | NASDAQ | DJIA |
|----------|------------|-----------|-----|--------|--------|------|
| ShangZheng | 0.2183 | 0.2183 | 0.3185 | 0.3589 | 0.3232 |
| ShenCheng | 0.2183 | 0.3185 | 0.2357 | 0.2047 | 0.2359 |
| HSI | 0.3185 | 0.3589 | 0.2357 | 0.1500 | 0.1731 |
| S&P500 | 0.3589 | 0.3232 | 0.2357 | 0.1500 | 0.2038 |
| NASDAQ | 0.3232 | 0.3232 | 0.2597 | 0.1731 | 0.2038 |
| DJIA | 0.3589 | 0.3589 | 0.2357 | 0.1500 | 0.2038 |

Table 3
Distances between all these time series from the year 1999 to the year 2006.

| Distance | ShangZheng | ShenCheng | HSI | S&P500 | NASDAQ | DJIA |
|----------|------------|-----------|-----|--------|--------|------|
| ShangZheng | 0.2014 | 0.2014 | 0.3300 | 0.3330 | 0.2550 |
| ShenCheng | 0.2014 | 0.3239 | 0.3092 | 0.2572 | 0.2580 |
| HSI | 0.3300 | 0.3239 | 0.3234 | 0.2958 | 0.2974 |
| S&P500 | 0.3330 | 0.3092 | 0.3234 | 0.2759 | 0.2207 |
| NASDAQ | 0.2550 | 0.2572 | 0.2958 | 0.2759 | 0.3545 |
| DJIA | 0.2956 | 0.2580 | 0.2974 | 0.2207 | 0.3545 |

Table 4
Distances between all these time series from the year 2007 to the year 2013.

| Distance | ShangZheng | ShenCheng | HSI | S&P500 | NASDAQ | DJIA |
|----------|------------|-----------|-----|--------|--------|------|
| ShangZheng | 0.2217 | 0.2217 | 0.3025 | 0.2538 | 0.2246 |
| ShenCheng | 0.2217 | 0.3208 | 0.3060 | 0.2678 | 0.3182 |
| HSI | 0.3025 | 0.3208 | 0.3382 | 0.3685 | 0.3377 |
| S&P500 | 0.2538 | 0.3060 | 0.3382 | 0.1557 | 0.1611 |
| NASDAQ | 0.2246 | 0.2678 | 0.3685 | 0.1557 | 0.2046 |
| DJIA | 0.2359 | 0.3182 | 0.3377 | 0.1611 | 0.2046 |

Stock time series are not symbolic sequences; therefore, it is to map the continuous variable of stock closing prices to a set of symbols. Let us consider a binary mapping rule, to map each pair of successive stock closing prices to the symbols 0 and 1, corresponding to a decrease or increase in the stock closing prices, respectively. This mapping rule has the advantage of being simple enough for practical purposes. So we turn Data A, B, C, D to symbolic sequences.

First, we calculate the dissimilarity index (we call 'distance' below) defined in Eq. (2) for every two symbolic time series in the whole stock time series from 1991 to 2013 (Data A). For each pair of time series in Data A, the distance with \( m = 6 \) is applied to the symbolic time series (Table 1). Then, we divide the whole time series into three subsets (Data B, C, D) and measure the distances between each pair of subjects who belong to the same subsets similarly (Tables 2–4). Greater distance indicates less similarity and vice versa. Our results show that the distances in Table 1 are almost smaller than the distances in Tables 2–4. Also, for every subset, we find that the distances between DJIA, NASDAQ, and S&P500 are quite close with small differences; the same results can be found for the distances of ShangZheng, ShenCheng, and HSI. Because the business behavior of the stock markets in the same area are influenced by similar rules and the mutual influence between markets. Besides, the US stock market is one of the world’s most developed stock market, it is a mature stock market. The three major US stock indices are S&P500, NASDAQ, and DJIA, they reflect the whole financial market trends in the United States to a certain extent, so the similar characters of these three stock markets may be more obvious. But the distances shown in Table 3 are quite bigger than those shown in Tables 1, 2 and 4. First, the result in Table 1 reflects characters of these three stock markets in the whole time period, it shows that the three stock indices have high auto-correlation on the macro perspective; but Table 3 reflects the result after the Asian currency crisis and before the global financial crisis, in this period, most areas of Asia were influenced by the crisis, leading to many disinvestment from western countries, and the markets became more liberal and liquid, the instability of the world economy market make the US stock market more complex and changeable. And the distances in the US markets or Chinese markets are smaller than the distances from inter markets. This implies higher similarity between the stock time series in the US stock markets or Chinese stock markets, but weaker
Fig. 2. The distance between (a) ShangZheng and the others, (b) ShenCheng and the others, (c) HSI and the others, (d) S&P500 and the others, (e) NASDAQ and the others, and (f) DJIA and the others in Data B, Data C, and Data D, respectively.

similarity in the inter markets. We know that the economic factors of stock indices in the same area should be similar, but for the Chinese stock market and the US stock market, they belong to two different countries, which have different economic background and political factors, so the similarity between two stock indices from two different stock markets should be weaker than from the same stock market. It means that in the same market, the similarity can be higher and have stronger auto-correlations.

In order to analyze the similarity between stock time series in different time periods. The distances of Data B, Data C or Data D for every two symbolic time series are shown in Fig. 2. Fig. 2(a)–(c) shows the distances of ShangZheng (SZ), ShenCheng (SC), and HSI stock indices with the others, respectively. It is easy to find that the distances between ShangZheng (SZ) and ShenCheng (SC) indices in 1991–1998 are greater than in 1999–2007 and 2007–2013, which implies that the similarity of the two Chinese stock indices becomes larger after the Asian currency crisis or the global financial crisis. Likewise, the similarity between the two Chinese stock indices and the three US stock markets are much larger after the two crises. But for the HSI index, the similarity with the US stock markets becomes weaker after the two crises. Besides, the distances between HSI index and ShangZheng become smaller after the two crises, but the distances between HSI and ShenCheng are almost the same. In early 1998, the American stock market was affected by the Asian currency crisis and the yen continued to fall, then the HangSheng index began to fall sharply. But it is the early time that Hong Kong has just return to China, its economic was still influenced by the western countries. So the HSI is still close to US stock market after the Asian currency crisis. Although the United States was the source of the financial crisis, HSI index had built closer correlation with Chinese stock market at that time. During the global financial crisis, the Hong Kong stock market was affected greatly, but recovered quickly with China’s economic policy; so the similarity of HSI with US stock market became weaker than with Chinese stock market after the global financial crisis. The similarity among the Chinese stock markets does not change significantly for the three time periods, but the similarity between the Chinese stock markets and the US stock markets changes significantly. Fig. 2(d)–(f) shows the distances of S&P500, NASDAQ, and DJIA stock indices with the others, respectively. We can find that the distances of the three US stock markets with ShangZheng (SZ) and ShenCheng (SC) indices become smaller after the Asian currency crisis or the global financial crisis, which implies that the similarity between the three US stock markets and ShangZheng (SZ) and ShenCheng (SC) indices are larger after the two crises, just opposite to the three US stock markets with HSI. For the mainland of Chinese stock market, the correlation between them do not have obvious change, the market was not mature and less affected by the Asian currency crisis; besides, although the market was affected heavily during the global financial crisis, with government’s policy, the market recovered gradually, so the similarity between them have little change. With the integration of China and the world economy, the Chinese stock
market was affected by the two crises, but has more correlation with the US stock market. Especially in the global financial crisis, the economic cooperation between China and western countries enhances the link between Chinese stock market and US stock market. Besides, in the US stock markets, the distances of one stock time series with two others become larger after the Asian currency crisis and smaller after the global financial crisis, which shows that the similarity among the three US stock markets become weaker after the Asian currency crisis and larger after the global financial crisis. We can find that the effect on the US stock market from the global financial crisis is heavier than from the Asian currency crisis, because the global financial crisis was worldwide and affect almost every US stock index, but the stock market became more stochastic after the Asian currency crisis; so the correlation between them become closer in the global financial crisis.

After all the pairwise distances (as measured by the dissimilarity indices) are obtained for all datasets, we can use some standard techniques of categorization to present our results. Here, we use the phylogenetic tree algorithm as an example. The method for constructing phylogenetic trees [30–32] is a useful tool to present our results since the algorithm arranges different groups on a branching tree to best fit the pairwise distance measurements. We note that the structure of the tree is consistent with the underlying complexity: and further down the branch the dynamics are more complex. The distance between any two groups is the summation of the horizontal lengths along the shortest path on the tree that connects them. In Figs. 3–6, we show the result of a rooted tree for the similarity between stock time series of \( m = 6 \). We can clearly find the results shown above from the phylogenetic tree, especially the abnormal changes for the HSI index from Figs. 5 and 6 after the Asian currency crisis or the global financial crisis. This phenomenon may partly be due to the HSI market that can easily be influenced by the external factors. Being different from ShangZheng and ShenCheng, which are under a strong government’s macrominor economy control, HSI is easily influenced by the Asian currency crisis or the global financial crisis and hard to return.

Also, it is obvious that the three US stock indices belong to the same branch except Fig. 4, similarly for the ShangZheng and ShenCheng indices. Besides, after the two big crises, the two Chinese stock indices and the three US stock indices belong to one big branch. It is consistent with the result from Fig. 2. It means that the Chinese stock markets become closer to the US stock markets after the crisis.
Fig. 5. Phylogenetic tree generated according to the distances between six stock indices from 1999 to 2006.

Fig. 6. Phylogenetic tree generated according to the distances between six stock indices from 2007 to 2013.

Table 5
The list of 10 stock indices.

| Area     | Index       |
|----------|-------------|
| Asia     | 1 ShangZheng |
|          | 2 ShenCheng |
|          | 3 HSI       |
|          | 4 Nikkei225(Japan) |
| America  | 5 S&P500    |
|          | 6 NASDAQ    |
|          | 7 DJIA      |
| Europe   | 8 DAX(Germany) |
|          | 9 CAC40(France) |
|          | 10 FTSE(UK) |

3.2. Analysis of financial time series in different areas

In this section, we use the information categorization method to analyze the daily records of 10 stock indices listed in Table 5 during the period 1991–2013. Like in Section 3.1, we map the 10 stock time series to symbolic time series. Then, we calculate the distance between the symbolic time series. Fig. 7 shows a phylogenetic tree of the 10 stock indices for $m = 8$; however, similar results are obtained for $m = 4–12$.

The structure of the tree reveals that the Nikkei 225 index and the HSI index are arranged along a separate branch clearly distinguishable from other stock indices, which implies that the similarity of HSI index with the other two Chinese stock indices is weaker than with the US stock markets and Europe stock markets. But the Japan stock market shows rare similarity with others. It means that Japan stock market has its unique property which is unlike other stock markets in different areas. The Nikkei 225 is the most widely quoted average of Japanese equities and an internationally recognized index. Japan is a developed country in the Asian area and not like the development of Chinese market; it has its independent economic operation system and has long history in economic market. As a mature stock market, Japan stock market should be unlike other stock markets in the Asian area. Though Japan has close trade with western countries, it is still affected by the
Fig. 7. Phylogenetic tree generated according to the distances between ten stock indices from 1991 to 2013. Each stock index has been color-coded according to the area which belongs to. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

geographical environment and policy of the Asian area, especially it increases the economic correlation with China in recent years, its stock market is different from the market in America and Europe area. So the Nikkei 225 is on a separate branch. Similarly, the three US stock indices and DAX stock index are arranged along another branch. Besides, the other two stock indices in Europe and the two Chinese stock indices belong to another two separate branches. All the stock indices in the same area show larger similarity than in a different area. The information categorization analysis is able to distinguish each one from the others. The result is consistent with the consequence in Ref. [16].

4. Conclusions

In this paper, we detect the similarity of the daily stock markets by investigating the distances of stock time series and measure the dissimilarity indices between them. The information categorization method consists of the rank-frequency method and the distance calculation. The rank-frequency method does not require that the length of the time series should be the same, which is convenient for calculation unlike other methods. The new distance measure, defined in Eq. (2), incorporates both a probabilistic weighting factor given by Shannon’s entropy and a term related to the number of words. Besides, the values of $m$ have little effect on the result, but when we calculate the distances of time series, we should ensure the numbers of different $m$-bit words are the same. So we should choose proper size of the $m$-bit word considering different conditions.

Based on this method, we analyze the similarity between the whole stock time series, compared with the similarity between the stock time series in three different time periods (1991–1998, 1998–2006, 2007–2013). We find that the Chinese stock markets show higher similarity with the US stock markets after the Asian currency crisis and the global financial crisis except for the HSI index in Hong Kong. But the similarity among the three US stock markets becomes weaker after the Asian currency crisis but larger after the global financial crisis. Besides, the similarity of the Hong Kong stock market with the Chinese stock markets is larger than with the US stock markets. Hence, one can conclude that the influence of the two big crises on the Chinese stock markets is different from the US stock markets. Then, we extend the number of analyzed stock indices by using the method. We acquire satisfactory and obvious result for the stock indices in Asia, US, and Europe stock markets. It provides a good choice to classify the different stock time series and analyze the similarity between them.

In summary, we introduce a quantitative measurement of dissimilarity among symbolic sequences. This derivation is based on the generic statistical physics assumptions and, therefore, can be applied to a wide range of problems. With the simple measure of similarity, we can categorize different types of symbolic sequences by using standard clustering algorithms. This classification of symbolic sequences may provide very useful information about the underlying dynamical processes that generate these sequences. Furthermore, we can change the way mapping the time series to symbolic sequences, it may provide another result for the research. Maybe the definition of the dissimilarity indices are different for different objects of study. When we use the method, we should make full analysis of the background of time series. The method has been widely used in the physiologic systems and literature. This information categorization method is potentially useful because of its ability to take into account both macroscopic structures and the microscopic details of the dynamics. Now we make a preliminary research by using this method, expecting that the method can be applied to do further study on stock markets, transport system, or foreign exchange markets.

Acknowledgments

The financial supports from the funds of the China National Science (Nos. 61071142 and 61371130) and the Beijing National Science (No. 4122059) are gratefully acknowledged.
References

[1] Rosario N. Mantegna, H. Eugene Stanley, Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, 2000.

[2] Rosario N. Mantegna, H. Eugene Stanley, Scaling behaviour in the dynamics of an economic index, Nature 376 (6535) (1995) 46–49.

[3] H. Eugene Stanley, The fragility of interdependence: Coupled networks switching phenomena, Bull. Amer. Phys. Soc. 58 (2013).

[4] Steven M. Pincus, Approximate entropy as a measure of system complexity, Proc. Natl. Acad. Sci. 88 (6) (1991) 2297–2301.

[5] Steven M. Pincus, Approximate entropy (ApEn) as a complexity measure, Chaos 5 (1) (1995) 110–117.

[6] Steven M. Pincus. [14] Quantifying complexity and regularity of neurobiological systems, Quant. Neuroendocrinol. 28 (1995) 136.

[7] Steven M. Pincus, Richard R. Viscarello, Approximate entropy: a regularity measure for fetal heart rate analysis, Obstet. Gynecol. 79 (2) (1992) 249–255.

[8] Stephanie A. Caswell Schuckers, Use of approximate entropy measurements to classify ventricular tachycardia and fibrillation, J. Electrocardiol. 31 (1998) 101–105.

[9] Joshua S. Richman, J. Randall Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol.-Heart Circulatory Physiol. 278 (6) (2000) H2039–H2049.

[10] Douglas E. Lake, Joshua S. Richman, M. Pamela Griffin, J. Randall Moorman, Sample entropy analysis of neonatal heart rate variability, Am. J. Physiol.-Regul. Integr. Comp. Physiol. 283 (3) (2002) R789–R797.

[11] Li-Zhi Liu, Xi-Yuan Qian, Heng-Yao Lu, Cross-sample entropy of foreign exchange time series, Physica A 389 (21) (2010) 4785–4792.

[12] Madalena Costa, Ary L. Goldberger, C.-K. Peng, Multiscale entropy analysis of complex physiologic time series, Phys. Rev. Lett. 89 (6) (2002) 068102.

[13] Madalena Costa, Ary L. Goldberger, C.-K. Peng, Multiscale entropy analysis of biological signals, Phys. Rev. E 71 (2) (2005) 021906.

[14] Ranjit A. Tharaisinghma, Georg A. Gottwald, On multiscale entropy analysis for physiological data, Physica A 366 (2006) 323–332.

[15] Jianan Xia, Pengjian Shang, Multiscale entropy analysis of financial time series, Fluct. Noise Lett. 11 (04) (2012).

[16] Jianan Xia, Pengjian Shang, Jing Wang, Wenbin Shi, Classifying of financial time series based on multiscale entropy and multiscale time irreversibility, Physica A (2014).

[17] Yi Yin, Pengjian Shang, Modified dfa and dcca approach for quantifying the multiscale correlation structure of financial markets, Physica A 392 (24) (2013) 6442–6457.

[18] Saeed Aghabozorgi, Ying Wah Teh, Stock market co-movement assessment using a three-phase clustering method, Expert Syst. Appl. 41 (4) (2014) 1301–1314.

[19] Anna Czpakiewicz, Pawel Majdosh, Grouping stock markets with time-varying copula-GARCH model, Finance a Uver: Czech J. Econ. Finance 64 (2) (2014).

[20] Albert C.-C. Yang, Shu-Shya Hsu, Huey-Wen Yien, Ary L. Goldberger, C.-K. Peng, Linguistic analysis of the human heartbeat using frequency and rank order statistics, Phys. Rev. Lett. 90 (10) (2003) 108103.

[21] Albert C.-C. Yang, Shu-Shya Hsu, Huey-Wen Yien, Ary L. Goldberger, C.-K. Peng, Yang et al. reply, Phys. Rev. Lett. 92 (10) (2004) 109802.

[22] T.H. Reijmers, Ron Wehrens, F.D. Daeyaert, P.J. Lewi, Lucgarde M.C. Buydens, Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences, Biosystems 49 (1) (1999) 31–43.

[23] Claude E. Shannon, A note on the concept of entropy, Bell Syst. Tech. J. 27 (1948) 379–423.

[24] Albert C.-C. Yang, C.-K. Peng, H.-W. Yien, Ary L. Goldberger, Information categorization approach to literary authorship disputes, Physica A 329 (3) (2003) 473–483.

[25] C.-K. Peng, Albert C.-C. Yang, Ary L. Goldberger, Statistical physics approach to categorize biologic signals: from heart rate dynamics to dna sequences, Chaos 17 (1) (2007) 015115.

[26] Ary L. Goldberger, C.-K. Peng, Genomic classification using an information-based similarity index: application to the SARS coronavirus, J. Comput. Biol. 12 (8) (2005) 1103–1116.

[27] Jia-Rong Yeh, Tzu-Yu Lin, Jiann-Shing Shieh, Yun Chen, Chung-Kang Peng, A novel blocking index based on similarity measurement applied in distinguishing the patterns of blood pressure signals at dynamically transitional situation, Biomed. Eng.: Appl. Basis Commun. 20 (02) (2008) 107–114.

[28] Yosef Ashkenazy, Plamen Ch. Ivanov, Shlomo Havlin, Chung-K. Peng, Ary L. Goldberger, H. Eugene Stanley, Magnitude and sign correlations in heartbeat fluctuations, Phys. Rev. Lett. 86 (9) (2001) 1900.

[29] http://finance.yahoo.com.

[30] J. Felsenstein, (PHYLIPI): Phylogenetic Inference Package, Version 3.5 c. 1993.

[31] Naruya Saitou, Masatoshi Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (4) (1987) 406–425.

[32] Roderick D.M. Page, Edward C. Holmes, Molecular Evolution: A Phylogenetic Approach, John Wiley & Sons, 2009.