Physica A 419 (2015) 651–658

Contents lists available at ScienceDirect

Physica A

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

Effect of extreme data loss on heart rate signals quantified by entropy analysis

Li Yu [a], Wang Jun [b], Li Jin [a,*], Liu Dazhao [a]

[a] College of Physics and Information Technology, Shaanxi Normal University, Xi'an 710062, China
[b] Image Processing and Image Communications Key Lab, Nanjing University of Posts & Telecommunications, Nanjing 210003, China

ARTICLE INFO

Article history:
Available online 7 August 2014

Keywords:
Data loss
Heart rate variability
Base-scale entropy
Approximate entropy

ABSTRACT

The phenomenon of data loss always occurs in the analysis of large databases. Maintaining the stability of analysis results in the event of data loss is very important. In this paper, we used a segmentation approach to generate a synthetic signal that is randomly wiped from data according to the Gaussian distribution and the exponential distribution of the original signal. Then, the logistic map is used as verification. Finally, two methods of measuring entropy—base-scale entropy and approximate entropy—are comparatively analyzed. Our results show the following: (1) Two key parameters—the percentage and the average length of removed data segments—can change the sequence complexity according to logistic map testing. (2) The calculation results have preferable stability for base-scale entropy analysis, which is not sensitive to data loss. (3) The loss percentage of HRV signals should be controlled below the range ($p = 30\%$), which can provide useful information in clinical applications.

1. Introduction

Currently, with the rapid development of information technology, the study of signal processing methods to concentrate, extract and purify information from large quantities of chaotic data is the current focus of theory and application research in many fields, including scientific research, computer simulation, and finance [1–3]. On this issue, experts all over the world have performed extensive research and exploration with resulting great achievements [1–7]. However, data loss or data uselessness is inevitably encountered in real-world signal analysis, which causes many difficulties in data analysis and application, especially in the census, environmental monitoring, medical science and other large longitudinal studies. For example, heart rate variability (HRV) signals, which indicate the change in instantaneous heart rate, are an important index in evaluating the function of the autonomic nervous system. The HRV signal acquisition process can be affected by many factors, such as, the complex operation of equipment, long-term monitoring, and loose electrode contacts. Thus, it is easy to cause partial data loss. Moreover, the acquired original signals often contain some non-normal data, such as noise signals, artifacts and ectopic pacemaker signals. When these signals are pretreated by certain calculation methods, a small error can lead to normal data loss.

Based on earlier research on the problem of data loss [8–10], for instance, Z. Chen et al. have studied the effect of removing fixed length segments from a signal and stitching together the remaining parts using the DFA correlation analysis method [8]. N.E. Romero et al. have applied four treatment methods of data loss: listwise deletion, pairwise deletion, mean imputation and multiple imputation, to analyze the effect of data loss on the tests for the presence of market mechanisms [9]. The effect

* Corresponding author.
E-mail address: lijin1997@gmail.com (J. Li).

http://dx.doi.org/10.1016/j.physa.2014.06.074

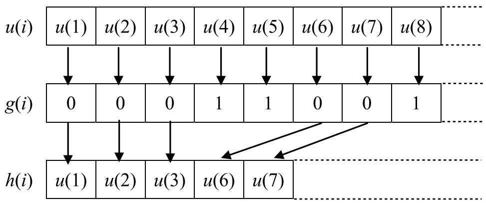

Fig. 1. Illustration of generating the synthetic signal $h(i)$ by removing data points from the intact original signal $u(i)$ according to a binary series $g(i)$.

of the percentage and the average length of removed data segments have been studied using the DFA scaling method by Qianli D.Y. Ma et al. [10]. We intend to focus more on the problem of data loss and explore more actively its effect on the estimation of multi-scale linear and nonlinear scaling measures. Meanwhile, when measuring whether a data processing method is reliable, we should pay more attention to the problem of whether the analysis results are stable, regardless of whether the original data are lost. Then, the hypothesis can be proposed that if an analysis method can maintain the stability of results in the situation of data loss, this method can analyze real-world data and predict the future trend according to the present state.

In this paper, the hypothesis is verified by the base-scale entropy method using a segmentation approach developed by Qianli D.Y. Ma et al. [10], which randomly wipes away data according to the probability distribution given by the original data. First, the logistic map is used as a verification. Then, two methods of entropy measurement—base-scale entropy and approximate entropy—were comparatively analyzed. The result showed that the approximate entropy (ApEn) is too sensitive to data loss for use in real-world signal analysis. However, the base-scale entropy has a stable result regardless of data loss and can thus be used to analyze real-world signals. Finally, to rule out the contingency of results through only one test, we calculated the respective base-scale entropy in 50 tests of data loss. The above results were verified, and we also found that the base-scale entropy can provide useful information for accurately judging physiological and pathological status as long as data loss is controlled below a certain range.

2. Methods

2.1. Synthetic signal generation

To simulate data loss, we first generate a binary series $g(i)$ with the same length N as the intact original series $u(i)$, using the segmentation approach [10]. In this method the positions i where $g(i) = 1$ will correspond to the positions at which data points in $u(i)$ are removed, while the positions where $g(i) = 0$ will correspond to the positions at which data points in $u(i)$ are preserved (Fig. 1) [10]. Finally, we generate a synthetic signal $h(i)$ by randomly removing data points from the original signal $u(i)$ and stitching together the remaining parts of $u(i)$ according to the binary series $g(i)$. The synthetic signals $h(i)$ are characterized by three parameters: (i) the percentage p of removed data, (ii) the average length u of the removed data segments, and (iii) the functional form $P(L)$ of the distribution of the length L of the removed data segments.

We used the following method to generate the binary series $g(i)$ [10]:

(i) We generate the lengths $L(j)$ ($j = 1, 2, 3, \ldots, M$) of the data segments to be removed from the intact original signal $u(i)$ by randomly drawing integer numbers from a given probability distribution $P(L)$ with mean value u. Each integer number drawn from $P(L)$ represents the length of a data segment removed from $u(i)$. The process continues until the summation of the lengths of all removed data segments equals or exceeds a predetermined amount $pN$of data to be removed, i.e.,

$$\sum_{j=1}^{M} L(j) \geq pN \tag{1}$$

where M is the minimal number to fulfill Eq. (1) [10]. Finally, we calculate the size of the last segment to obtain the exact fraction pN of the lost data.

(ii) We append a "0" to each element in the series $\{L(j)\}$ to act as a separator between two adjacent segments [see step(iv)], resulting in a new series $\{[L(j), 0]\}$. Note that now the sum of the series is $pN + M$.

(iii) We append $[N - (pN + M)]$ number of elements "0" to the end of the series $\{[L(j), 0]\}$ to make an extended series where the sum of all elements is N, equal to the length of the original series $u(i)$. Then, this extended series is shuffled, leading to a set of M elements $[L(j), 0]$ randomly decentralized in a "sea" of $[N - (pN + M)]$ elements "0" [see Eq. (2)] [10].

(iv) Next, we replace the numbers $L(j)$ in Eq. (2) with $L(j)$ number of elements "1" to obtain the binary series $g(i)$ as shown in Eq. (3) [10].

$$H(i) = \{\ldots, 0, [L(j), 0], 0, \ldots, 0, [L(j+1), 0], [L(j+2), 0], 0, \ldots\} \tag{2}$$

$$g(i) = \{\ldots, 0, \overbrace{\underbrace{1, \ldots, 1}, 0}, 0, \ldots, 0, \overbrace{\underbrace{1, \ldots, 1}, 0}, \overbrace{\underbrace{1, \ldots, 1}, 0}, 0, \ldots, \}. \tag{3}$$