Computers in Biology and Medicine 95 (2018) 13–23

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Appropriate use of the increment entropy for electrophysiological time series

Xiaofeng Liu [a,d,e,*], Xue Wang [b,c], Xu Zhou [a,d,e], Aimin Jiang [a,d,e]

[a] College of IoT Engineering, Hohai University, Changzhou 213022, China
[b] College of Computer and Information, Hohai University, Nanjing 210098, China
[c] School of Information and Engineering, Changzhou University, Changzhou 213164, China
[d] Changzhou Key Laboratory of Robotics and Intelligent Technology, Changzhou 213022, China
[e] Jiangsu Key Laboratory of Special Robots (Hohai University), Changzhou 213022, China

ARTICLE INFO

Keywords:
Increment entropy
Complexity
Electrophysiological time series

ABSTRACT

The increment entropy (IncrEn) is a new measure for quantifying the complexity of a time series. There are three critical parameters in the IncrEn calculation: N (length of the time series), m (dimensionality), and q (quantifying precision). However, the question of how to choose the most appropriate combination of IncrEn parameters for short datasets has not been extensively explored. The purpose of this research was to provide guidance on choosing suitable IncrEn parameters for short datasets by exploring the effects of varying the parameter values. We used simulated data, epileptic EEG data and cardiac interbeat (RR) data to investigate the effects of the parameters on the calculated IncrEn values. The results reveal that IncrEn is sensitive to changes in m, q and N for short datasets ($N \leq 500$). However, IncrEn reaches stability at a data length of $N = 1000$ with $m = 2$ and $q = 2$, and for short datasets ($N = 100$), it shows better relative consistency with $2 \leq m \leq 6$ and $2 \leq q \leq 8$ We suggest that the value of N should be no less than 100. To enable a clear distinction between different classes based on IncrEn, we recommend that m and q should take values between 2 and 4. With appropriate parameters, IncrEn enables the effective detection of complexity variations in physiological time series, suggesting that IncrEn should be useful for the analysis of physiological time series in clinical applications.

1. Introduction

In recent years, many scholars have attempted to derive the laws governing complex systems by using statistical approaches. Entropy is defined as the lack of information or information quantity when computing the probability distribution of the complexity of a time series or signal [1]. At present, entropy has been widely applied to analyse signals in various fields, such as medicine [2–6], finance [7,8] and ecology [9]. Over the past 27 years, entropy has been increasingly applied to analyse various physiological signals (Fig. 1), such as cardiac data and epilepsy data. Entropy has been used to describe the changes in cardiac signals under different physiological and pathological conditions [10–12] and to characterize epilepsy data for identifying [13–19] or predicting [20–22] seizures. Various entropy measures have been established over the past two decades; the approximate entropy (ApEn), sample entropy (SampEn) and permutation entropy (PE) are the most commonly used measures for analysing physiological time series. Fig. 1 shows that entropy was not used in the analysis of physiological signals before the 1990s because the calculation of early entropy measures often required a large amount of data [23,24], and gathering sufficient data was typically difficult or impossible in the case of physiological time series because of the effects of disease or age. In 1991, Pincus proposed the ApEn measure, which can be used to calculate the complexity of finite datasets and even short datasets [23]. ApEn measures the frequency of similar epochs in a time series; more frequent and more similar results lower the ApEn value. Researchers have used ApEn to extract the features of different stages of epileptic time series [25,26], and the results demonstrate that the ApEn value decreases from the interictal stage to the ictal stage. However, ApEn also has some disadvantages: it shows an inherent bias towards regularity due to self-matching, there is a lack of relative consistency among ApEn values calculated with different combinations of parameters, and it is sensitive to the dataset length [23,24]. SampEn is similar to ApEn, but SampEn overcomes the shortcomings of ApEn: it avoids the self-matching of vectors, shows good relative

* Corresponding author. College of IoT Engineering, Hohai University, Changzhou 213022, China.
E-mail addresses: xfliu@hhu.edu.cn (X. Liu), wangxue@cczu.edu.cn (X. Wang), hhuc_zx@hhu.edu.cn (X. Zhou), jiangam@hhuc.edu.cn (A. Jiang).

https://doi.org/10.1016/j.compbiomed.2018.01.009
Received 31 July 2017; Received in revised form 23 January 2018; Accepted 23 January 2018

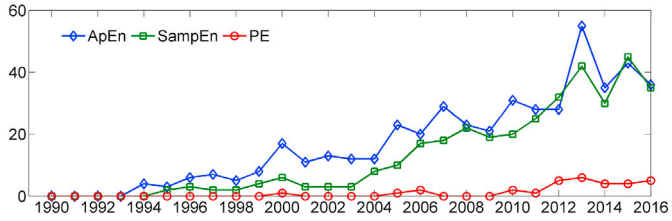

Fig. 1. Total number of publications considering the entropy of physiological signals listed on the Web of Science from 1990 to 2016.

consistency and is independent of the dataset length [24]. SampEn exhibits better performance on biological signals, such as heart rate data, epilepsy data and gait data [27,28].

Although ApEn and SampEn are often used to analyse the complexity of electrophysiological time series, they both ignore the temporal order of the elements in a signal [29]. In 2002, Bandt and Pompe proposed PE, which is a symbolic dynamic measure based on the natural ordinal pattern of a time series [29]. Following its proposal, PE was widely adopted for use in epileptic seizure identification and prediction [30–34]. Xiaoli Li et al. [34] compared PE and SampEn in absence seizure. They found that PE is better able to predict absence seizures. However, PE considers only the order of the values in a time series; it ignores the changes in magnitude between the elements of the time series. Subsequently, by incorporating magnitude information into the mapped patterns, many variations of PE have been developed, such as the fine-grained permutation entropy [35] and the weighted-permutation entropy [36]. The fine-grained permutation entropy is sensitive to abrupt changes [32,37]; however, the weighted-permutation entropy improves on this problem. In addition, in the case of equal values, the processing method for PE ignores equal values or treats all equal values as one symbol [38]. Therefore, a question arises as to how to consider the lengths of sequences of adjacent data in a time series. The increment entropy (IncrEn), which was proposed by Liu et al. [39], quantifies the magnitudes of the variations between adjacent elements into ranks based on a precision factor and the standard deviation of the time series. Liu and his colleagues demonstrated that IncrEn shows better performance for seizure detection than either PE or SampEn and better discrimination when there are subtle changes in structure or energy in a time series.

IncrEn is conceptually similar to PE in that it also uses the concepts of entropy and symbolic dynamics. In the IncrEn calculation, two letters are used to describe the relationship between adjacent elements in a time series. One letter represents the volatility direction, and the other represents the magnitude of the variation between the adjacent elements. In this approach, a raw time series is reconstructed into many vectors, each consisting of m elements. Each element of each vector represents the increment between two neighbouring elements in the original time series. Each increment element is mapped to a word consisting of two letters, and then, each vector is described in terms of a symbol sequence pattern. In the IncrEn approach, the complexity is evaluated by using the Shannon entropy to calculate the probabilities of independent patterns. A larger number of independent patterns corresponds to a higher IncrEn value. There are three parameters in the IncrEn calculation: the length of the time series (N), the dimensionality (m) and the quantifying precision (q). m represents the window size, or the length of the vectors that are considered in the comparison, whereas q represents the precision of the fluctuation amplitudes. Liu and his colleagues estimated the identification capability of IncrEn on many kinds of signals, such as simulated data and intracranial epileptic signals. However, the dependence of the effectiveness and consistency of IncrEn on the choice of the parameters used for the analysis of physiological time series has not yet been investigated. If m is too large, it will be difficult to identify changes in the time series [39]. If q is too large, IncrEn will be too sensitive to noise [39]. Thus, the selection of an appropriate combination of these parameters based on the corresponding characteristics of the IncrEn measure is very important for achieving desirable identification results. Therefore, the aims of this paper are (1) to investigate the effects of the three parameters on IncrEn, (2) to characterize the relative consistency of IncrEn, (3) to investigate the ability of IncrEn to distinguish between two classes based on physiological time series, and (4) to test appropriate parameter combinations for the IncrEn calculation.

The remainder of the paper is organized as follows. The definition of IncrEn is introduced in Section 2, and Section 3 describes the calculation of IncrEn in detail. Section 4 introduces the datasets used to conduct our experiments and illustrates the influence of the parameters on IncrEn. Section 5 reports tests of appropriate parameter choices for the IncrEn calculation. Section 6 presents a discussion and conclusion.

2. Definition of IncrEn

The IncrEn approach is conceptually similar to the PE approach, which naturally encodes the rank order of a time series in the form of symbol sequences. However, the IncrEn calculation considers not only the volatility directions but also the magnitudes of variation between adjacent elements. For a time series $\{x(i), 1 \le i \le N\}$, where N is the data length, an increment series $\{v(i), 1 \le i \le N-1\}$ is constructed, where $v(i) = x(i+1) - x(i)$. IncrEn can then be calculated with a chosen number m of embedding dimensions. The increment series is divided into $N-m$ vectors, each of m dimensions. Each element in each vector is mapped to a word consisting of two letters, the sign and the size. The sign indicates the direction of the volatility between the corresponding neighbouring elements in the original time series; it takes values of 1, 0, or −1, indicating a rise, no change, or a decline, respectively. The size describes the magnitude of the variation between these adjacent elements. Thus, the original time series is mapped to $N-m$ words of $2m$ letters each. Let q be the quantifying precision of the variation between adjacent elements. Then, for a word of $2m$ letters, there are $(2q+1)^m$ possible distinct variations. Let w_n denote the nth unique word. $Q(w_n)$ is the total number of instances of the nth unique word; thus, we can define the relative frequency of each unique word as $P(w_n) = \frac{Q(w_n)}{N-m}$. Finally, IncrEn is defined as shown in Eq. (1).

$$H(m) = -\frac{1}{m-1}\sum_{n=1}^{(2q+1)^m} P(w_n)\log P(w_n). \quad (1)$$

here, $m-1$ is a normalization factor. $H(m)$ is bounded on $\left[0, \ \frac{m\log(2q+1)}{(m-1)}\right]$.

3. Calculation of IncrEn

In the IncrEn approach, the complexity of a time series is measured by encoding the signs and sizes of the variations between adjacent elements to reflect its natural fluctuations. The computation of IncrEn for a given time series $x\{i\}$ of length N consists of six steps [39]:

I Construct the increment time series $\{v(j), 1 \le j \le N-1\}$ from the original time series $\{x(i)\}$.

II Divide the increment time series $\{v(j)\}$ into vectors of m points in length. m must be fixed before the IncrEn calculation. This division will result in $N-m$ increment vectors, $V(l) = [v(l), \ldots, v(l+m-1)], 1 \le l \le N-m$.

III Calculate the pattern vector w_l for each increment vector $V(l)$. First, calculate the mapped word for each element, where the sign is $s_k = sgn(v(k))$ and the size q_k is equal to the minimum between q and $\left|\frac{v(j)\times q}{std(\{v(j)\})}\right|$ or to zero if $std(\{v(j)\}) = 0$. q must also be fixed before the IncrEn calculation. w_l is constructed as the combination of all corresponding s_k and q_k pairs.

IV Count the total number of instances $Q(w_n)$ of every unique word w_n that occurs in $\{v(j)\}$.