Pattern Recognition 46 (2013) 808–816

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Possibilistic nonlinear dynamical analysis for pattern recognition

Tuan D. Pham*

Aizu Research Cluster for Medical Engineering and Informatics, Center for Advanced Information Science and Technology, The University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan

ARTICLE INFO

Article history:
Received 15 January 2012
Received in revised form
28 August 2012
Accepted 30 September 2012
Available online 11 October 2012

Keywords:
Nonlinear dynamics
Entropy measures
Possibility
Fuzzy sets
Geostatistics
Biosignals
Pattern recognition

ABSTRACT

A nonlinear dynamical system can be defined as a study of any system that implies motion, change, or evolution in time where a change in one variable is not proportional to a change in a related variable. The mathematical operations underlying such a system are very useful for pattern recognition with time-series data. One of the most recent developments in nonlinear dynamical analysis is the so-called approximate entropy family. However, its algorithms are deterministic and do not consider uncertainty where the modeling of possibility can be appropriate and advantageous in many practical situations. Thus, possibilistic entropy algorithms are proposed in this paper as a new methodology for nonlinear dynamical analysis. The proposed approach is based on the notions of the approximate entropy family, geostatistics, and the theory of fuzzy sets. Furthermore, for the first time, nonlinear dynamical analysis of mass spectrometry data is presented for computer-based recognition of potential protein biomarkers and classification, which can be utilized for early disease prediction. Experimental results using proteomic and genetic data have shown the potential application of the proposed possibilistic nonlinear dynamical analysis to the study of complex biosignals.

1. Introduction

Nonlinear dynamical analysis methods derived from the information theory for measuring the complexity of time-series data have been successfully applied to many scientific disciplines, including biology, physiology, medicine, biophysics, chemistry, and economics [1]. In fact, analysis of similarity of time series is an active area of research in pattern classification and recognition [2–4]. However, the impact of these methods has only been partly explored to date for a better understanding of physiological function [5]. Some important physiological findings based on the concepts of nonlinear dynamics were also addressed in [5], including four major methodology families: fractals, entropy measures, symbolic dynamics measures, and Poincaré plot representation. Among these four families, the entropy measures are the most widely used methods for studying biological and physiological time-series data. In fact, the concept of entropy is variously defined in physics, mathematics, statistics, computer science, engineering, life science, economics, and many other disciplines. Because of its general characteristics, it has many interpretations and been a confusing idea to many researchers of different study fields [6,7]. John von Neumann actually suggested the term "entropy" because "no one knows what entropy really is, so in debate you will always have the advantage" [7].

After the introduction of the mathematical definition of entropy into the theory of information by Shannon [8], there have been several extensions of its principle. Popular types of entropy include fuzzy entropy [9], Kolmogorov–Sinai entropy [10,11], approximate entropy (ApEn) [12], etc. In particular, the fuzzy entropy defined in [9] replaces the probabilities of the values of a random variable by the fuzzy membership grades of a fuzzy set. It is therefore considered as a measure of the fuzziness of a fuzzy set. The entropy approach discussed in this paper refers to the original definition of the approximate entropy (ApEn), which was developed for understanding signal predictability or system complexity. The first method of this entropy family, known as approximate entropy (ApEn), was developed by Pincus [12–14]. ApEn is rooted in the work of Grassberger and Procaccia [15] and Eckmann and Ruelle [16], and widely applied in clinical cardiovascular studies and analysis of biomedical signals [17,18]. A low value of the approximate entropy indicates the time series is deterministic (low complexity), whereas a high value indicates the data is subject to randomness (high complexity) and therefore difficult to predict. In other words, lower entropy values indicate more regular the signals under study, whereas higher entropy values indicate more irregular the signals.

Extending the framework of approximate entropy (ApEn), sample entropy (SampEn) [19] and multiscale entropy (MSE) [20] was introduced to enhance the predictability analysis of

* Tel.: +81 242 37 3216.
E-mail address: tdpham@u-aizu.ac.jp

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.09.025

time-series data with particular reference to physiological signals. In general, both ApEn and SampEn estimate the probability that the sequences in a dataset which are initially closely related remain closely related, within a given tolerance, on the next incremental comparison. ApEn differs from SampEn in that its calculation involves counting a self-match for each sequence of a pattern, which leads to bias in ApEn [14]. SampEn is precisely the negative natural logarithm of the conditional probability that two sequences similar for m points remain similar at the next point, where self-matches are not included in calculation of the probability. Thus a lower value of SampEn also indicates more self-similarity in the time series. Based on the concept of fuzzy sets, a method named FuzzyEn was developed [21], where the similarity is defined by the degree of fuzziness and the shapes of the fuzzy membership functions.

This family of entropy measures has been increasingly applied to many problems in biomedical engineering and other fields of life sciences [22,23]. However, it has been pointed out that ApEn suffers from two major drawbacks: (1) because it is a function of the length of the sequence under study, it yields entropy values lower than expected for short sequences, being due to the counting of a self-match for each sequence, which leads to bias [14] and (2) it can be inconsistent with different testing conditions using different parameters of the entropy index. SampEn does not count self-matches and therefore can reduce bias. It has been found that SampEn can provide better relative consistency than ApEn because it is largely independent of sequence length [19]. MSE measures complexity of time-series data by taking into account multiple time scales, but MSE uses SampEn to quantify the regularity of the data. Most recently, as another entropy method, namely GeoEntropy (GeoEn), has been developed [24]. Although GeoEn can relax the assumption of the parameter selections encountered by other entropy-based methods, it does not allow the continuous modeling of the similarity measure. Based on the motivation that the theory of fuzzy sets has been found to be useful for analysis of complex physiological signals [21], two new possibilistic entropy methods, namely PossEnH and PossEnP, with particular reference to the study of biomedical signals, are introduced in this paper, which have the capability of identifying the correlated structural (spatial) information of mass spectrometry data, based on which multiple potential biomarkers can be selected. These entropy measures are based on the notion of the theory of possibility [25], which is a fuzzy restriction acting as an elastic constraint on the values that may be assigned to the variable of similarity in our study. The development of a possibilistic entropy using the ordinary kriging scheme (PossEnP) has been briefly reported in the literature [26]. This paper largely extends both technical and experimental discussions of the possibilistic entropy in the context of PossEnH and PossEnP, which can be useful for different purposes of applications.

The rest of this paper is organized as follows. Section 2 presents a possibilistic entropy measure as an extension of GeoEn by the modeling of possibility in uncertainty considered as an alternative to probability. Section 3 presents another possibilistic entropy measure which is based on the theory of possibility and a kriging estimator. Three experiments using three datasets to test the performance of the possibilistic entropy analysis of proteomic mass spectra of major adverse cardiac events for identifying potential biomarkers, cancer classification, and DNA similarity searching are discussed in Section 4. Finally, Section 5 is the conclusion of the research finding.

2. Possibilistic entropy: extended GeoEntropy

Let X_N be a time series of length $N: X_N=\{x_1,\ldots,x_N\}$ and Q_m be the set of all subsequences of the same length m in $X_N: Q_m=\{X_{1m},\ldots,X_{(N-m+1)m}\}$, where $X_{im}=\{x_i,\ldots,x_{i+m-1}\}$. It is said that X_{im} and X_{jm} are similar if and only if

$$|x_{i+k}-x_{j+k}|<r \quad \forall k,\ 0\le k<m \tag{1}$$

where r is threshold for similarity.

The probability of patterns of length m that are similar to the pattern of the same length that begins at i is

$$C_{im}(r)=\frac{K_{im}(r)}{N-m+1} \tag{2}$$

where $K_{im}(r)$ is the number of subsequences in Q_m that are similar to X_{im}.

The total average probability $C_{im}(r)$ for all i, $i=1,\ldots,N-m+1$, is

$$C_m(r)=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}C_{im}(r) \tag{3}$$

The approximate entropy (ApEn), given length m and tolerance value r, can now be readily computed by

$$ApEn(m,r)=\log\left[\frac{C_m(r)}{C_{m+1}(r)}\right] \tag{4}$$

To avoid bias in self-matching encountered in ApEn, sample entropy (SampEn) works in a slightly different way by defining X_{im} and X_{jm} are similar if and only if

$$|x_{i+k}-x_{j+k}|<r \quad \forall k,\ 0\le k<m,\ i\ne j \tag{5}$$

Let $L_m=\{X_{1m},\ldots,X_{(N-m-1)m}\}$ be the probability of patterns of length m that are similar to the pattern of the same length that begins at i is

$$B_{im}(r)=\frac{J_{im}(r)}{N-m-1} \tag{6}$$

where $J_{im}(r)$ is the number of subsequences in L_m that are similar to X_{im}.

The total average probability $B_{im}(r)$ for all i, $i=1,\ldots,N-m$, is

$$B_m(r)=\frac{1}{N-m}\sum_{i=1}^{N-m}B_{im}(r) \tag{7}$$

Finally, the value of SampEn, given m and r, can be calculated by the following equation:

$$SampEn(m,r)=\log\left[\frac{B_m(r)}{B_{m+1}(r)}\right] \tag{8}$$

Although approximate entropy has been widely used for studying the complexity of biosignals, it suffers from two major drawbacks [22]: (1) because it is a function of the length of the sequence under study, it yields values lower than expected for short sequences; (2) it can be inconsistent with different testing conditions using different parameters of the entropy index. SampEn does not count self-matches and therefore can reduce bias. It has been found that SampEn can provide better relative consistency than ApEn because it is largely independent of sequence length [19]. MSE measures complexity of time series data by taking into account multiple time scales, and uses SampEn to quantify the regularity of the data. All of these three methods depend on the selection of the two parameters known as m and r: parameter m is used to determine the sequence length, whereas parameter r is the tolerance threshold for computing pattern similarity. Results are sensitive to the selections of these two parameters and it has recently been reported that good estimates of these parameters for different types of signals are not easy to obtain [27]. GeoEn [24] introduces a geostatistical distance to provide a solution for the determination of m and r. Its principle is based on the theory of regionalized variables in geostatistics [28]. However, criterion of GeoEn for determining the similarity is based on a hard threshold as the absolute