KRIGING-BASED POSSIBILISTIC ENTROPY OF BIOSIGNALS

Tuan D. Pham

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

ABSTRACT

This paper presents an approach for nonlinear dynamical analysis of complex time-series data using the principles of the approximate entropy family, geostatistics, and possibility. Uncertainty of the measure of signal similarity is modeled using the concept of fuzzy sets and quantified by the signal error matching. The proposed method has the ability to discern the signal complexity at a more detailed level than the approximate entropy as well as to incorporate the spatial information inherently existing in the signal characteristics. Based on experimental results on the study of mass spectrometry data for cancer study, the proposed method appears to be a promising tool for classification of biosignals.

Index Terms— Nonlinear signal processing, approximate entropy, geostatistics, kriging, fuzzy sets, biosignals.

1. INTRODUCTION

After the introduction of the mathematical definition of entropy into the theory of information by Shannon [1], there have been several extensions of its principle. Popular types of entropy include fuzzy entropy [2], Kolmogorov-Sinai entropy [3], approximate entropy (ApEn) [4], etc. In particular, the fuzzy entropy defined in [2] 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. In fact, nonlinear signal analysis methods derived from the information theory have been successfully applied to many scientific disciplines, including biology, medicine, chemistry, and economics [5].

The entropy approach discussed in this paper refers to the framework 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 [4]. ApEn is rooted in the work of Grassberger and Procaccia [6] and Eckmann and Ruelle [7], and widely applied in clinical cardiovascular studies and analysis of biomedical signals [8, 9]. 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) [10] was introduced to enhance the predictability analysis of time-series data with particular reference to physiological signals. This family of entropy measures have been increasingly applied to many problems in biomedical engineering and other fields of life sciences [11]. In general, the approximate entropy famility is a methodology for studying nonlinear dynamical systems, which 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 is very useful for pattern recognition involving with time-series data. However, its algorithms are deterministic and do not consider uncertainty where the modeling of possibility can be appropriate and advantageous in many practical situations.

Based on the motivation that the theory of fuzzy sets has been found to be useful for analysis of complex physiological signals [12], a new possibilistic entropy method, with particular reference to the study of biomedical signals, is introduced in this paper, which have the capability of identifying the correlated structural (spatial) information of mass spectrometry data, based on which diseased and control samples can be better classified. This entropy measure is based on the notion of the theory of possibility [13], which is a fuzzy restriction acting as an elastic constraint on the values that may be assigned to the variable of similarity; and the derivation of kriging-based estimate error for matching signal similarity. The possibilistic entropy has recently been introduced in [14] to identify the cohorts of potential biomarkers from the mass spectrometry data of major adverse cardiac events. Here the method is further applied to the problem of cancer classification using mass spectrometry data.

The rest of this paper is organized as follows. Section 2 presents the proposed entropy using the principles of geostatsitcs and possibility. The kriging-based possibilistic entropy analysis of proteomic mass spectra for classifying cancer and control samples is discussed in Section 3. Finally, Section 4 is the conclusion of the research finding.

2. KRIGING-BASED POSSIBILISTIC ENTROPY

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

$$|x_{i+k} - x_{j+k}| < r, \forall k, 0 \leq k < m \quad (1)$$

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

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

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

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

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

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

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

$$|x_{i+k} - x_{j+k}| < r, \forall k, 0 \leq k < m, i \neq j \quad (5)$$

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

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

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

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

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

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

Having discussed ApEN and SampEn, we seek to present a model for quantifying signal predictability which is free from the specification of the critical parameter r and provides an analytical form for estimating the parameter m as follows.

Consider a function $\gamma(h)$, which is called an experimental semi-variogram of a sequence X and defined as [15]

$$\gamma(h) = \frac{1}{2(n-h)} \sum_{i=1}^{n-h} (x_i - x_{i+h})^2 \quad (9)$$

where x_i is a value of X taken at location i, x_{i+h} another value taken at h distance away (for $h = 1$ in a time-series signal, every point is compared with its neighbors; and for $h = 2$, every point is compared with a point two spaces away), and n is the total number of points which gives $(n - h)$ as the total number of the pairs of points.

In geostatistics, the semi-variograms are functions which describe the degree of spatial dependence of a spatial random field. It is defined as the variance of the difference between field values at two locations across the realizations of the field [16]. In addition to the experimental semi-variogram, there are basic or theoretical semi-variogram models which can be classified into two types: $\gamma(h)$ reach a plateau at some value of h, and $\gamma(h)$ does not reach a plateau. The most commonly used theoretical model is the spherical function, which is defined as [15]

$$\gamma(h) = \begin{cases} s \left[1.5\frac{h}{g} - 0.5(\frac{h}{g})^3 \right] & : \quad h \leq g \\ s & : \quad h > g \end{cases} \quad (10)$$

where g and s are called the *range* and *sill* of the theoretical semi-variogram, respectively; the first and third power of (h/g) are used to construct a curve being linearly increasing at small separation distances near the origin but becoming flattening at larger distances; and when s =1, it becomes the standardized function.

The above geostatistical terminology terms that are used to describe the important features of the spherical semi-variogram are explained as follows [15]. As the separation distance between data pairs increases, the corresponding semi-variogram value will also increase. However, an increase in the separation distance eventually no longer observes a corresponding increase in the averaged squared difference between the pairs of data, and the semi-variogram converges to a plateau. The distance at which the variogram reaches this plateau is called the *range* which can be specified as the value for the parameter m used by ApEn and SampEn.

Now let $x_z \in X_{im}$ be a value located at position z, $z \in [i, i+m-1]$. The value of x_z is supposed to be unknown and estimated using the following ordinary kriging system [15]:

$$\mathbf{C}\,\mathbf{a} = \mathbf{b} \tag{11}$$

where $\mathbf{C}$ is the square and symmetrical matrix that represents the spatial covariances between the known values $x_w \in X_{im}$, $w \neq z$, $w \in [i, i+m-1]$; and $\mathbf{b}$ is the vector that represents the spatial covariances between the x_z and x_w. To simplify the mathematical notation, let $(y_1, \ldots, y_p)$, $p < m$, be the sequence of p known values x_w, we have

$$\mathbf{C} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} & 1 \\ . & . & \cdots & . & . \\ . & . & \cdots & . & . \\ . & . & \cdots & . & . \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

where γ_{12} is the semi-variance of y_1 and y_2.

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_p & -\lambda \end{bmatrix}^T$$

where $a_1, \ldots, a_p$ are called the kriging weights, and λ is a Lagrange multiplier.

$$\mathbf{b} = \begin{bmatrix} \gamma_{z1} & \gamma_{z2} & \cdots & \gamma_{zp} & 1 \end{bmatrix}^T$$

where γ_{z1} is the semi-variance of x_z and y_1.

Thus the vector of the spatial predictor coefficients can be obtained by solving: $\mathbf{a} = \mathbf{C}^{-1}\,\mathbf{b}$, and the minimized kriging error variance is given by

$$\sigma_E^2 = \mathbf{a}^T \mathbf{b}. \tag{12}$$

We consider the notion of distortion measures between two feature vectors in signal processing, where the mismatch/dissimilarity between the two vectors can be quantified by a distortion using their predictor coefficients. Intuitively, a match of the two vectors is good if the distortion is small. A popular distortion measure is the error ratio-based distortion [17], where the predictor coefficients of one vector is used to estimate the prediction error produced by the other vector. If the two vectors are identical then both prediction errors should be the same, hence the distortion is zero (by using 1 minus the error ratio which is 1); otherwise a distortion value exists. Based on the error ratio-based distortion framework, we can derive a fuzzy membership function for quantifying the possibility of similarity between X_{im} and X_{jm} in terms of the kriging error variances of the two subsequences as

$$u_{ij}(p) = \frac{\mathbf{a}_i^T \mathbf{b}_i}{\mathbf{a}_j^T \mathbf{b}_i} \tag{13}$$

where $\mathbf{a}_i$ is defined in Eq. (11) which is the kriging prediction vector of X_{im}, $\mathbf{b}_i$ is defined in Eq. (11) associated with X_{im} , and $\mathbf{a}_j$ is the kriging prediction vector of X_{jm}. It can be seen that the use of Eq. (13) allows a convenient way for processing long signals by considering only the prediction coefficients. The possibility of signal similarity is embedded in the ratio of the distortion in a continuous scale. When the two subsequences are identical, their degree of similarity has its maximum value of 1; otherwise, the degree of similarity decreases according to the increasing mismatch of the two subsequences.

The possibility, instead of probability, of a subsequence X_{im} being similar to all other subsequences X_{jm} can be defined as

$$\omega_i^m(p) = \frac{1}{N-m-1} \sum_{j=1, j \neq i}^{N-m} u_{ij}(p) \tag{14}$$

The possibility, instead of probability, of all sub-sequences X_{im} being similar to all other subsequences X_{jm} is given by

$$\phi^m(p) = \frac{1}{N-m} \sum_{i=1}^{N-m} \omega_i^m(p) \tag{15}$$

Following the definition of SampEn expressed in Eq. (8), the possibilistic entropy PossEnP can be defined as

$$PossEnP = \ln \phi^m(p) - \ln \phi^{m+1}(p) \tag{16}$$

Procedure for calculation of PossEnP:

1. Given a finite signal $X_N = [x_1, x_2, \ldots, x_N]$, and set p.
2. Compute the expermental semi-variance of X_N using Eq. (9), and estimate its range $g(X_N)$ by fitting Eq. (9) into Eq. (10).
3. Set vector length $m = g(X_N)$.
4. Construct vectors of length m, X_{1m} to $X_{(N-m)m}$, defined as

   $$X_{im} = (x_i, x_{i+1}, \ldots, x_{i+m-1}),\ 1 \leq i \leq N-m$$

5. Estimate the midpoints of X_{im} and X_{jm} using Eq. (11).
6. Compute $u_{ij}(p)$, the possibility of similarity of X_{im} and X_{jm}, using Eq. (13).
7. Compute $\omega_i^m(p)$, the total possibility of similarity of X_{im} and X_{jm} using Eq. (14).
8. Compute $\phi^m(p)$, the total average possibility of similarity for all X_{im}, using Eq. (15).
9. Set $m = m+1$ and repeat steps 4-8 to obtain $\phi^{m+1}(p)$.
10. Calculate PossEnP, the possibilistic entropy of X_N at p, using Eq. (16).

Fig. 1. A typical mass spectrum (ovarian cancer)

3. ANALYSIS OF COMPLEX BIOSIGNALS

A mass spectrum is an x-y plot of intensity against mass-to-charge (m/z) ratio of a separated chemical collection. The mass spectrum of a given sample is the distribution pattern of the components of that collection based their mass-charge ratio. The m/z ratio is obtained by dividing the mass number of an ion by its charge number. For mass analysers such as time of flight, the direct x-axis measurement is the time series of the ions measured by the detector. The y-axis of a mass spectrum represents the signal intensity of the ions, and has arbitrary units. Figure 1 shows a mass spectral signal of ovarian cancer obtained from the FDA-NCI Clinical Proteomics Program Databank. The distinctive indicators of a proteomic mass-spectrum are referred to as biomarkers, which can be used to improve the efficiency of drug discovery and development. Mass spectrometry has been known as an important analytical technique for biomarker discovery and evaluation. The important role of mass spectrometry is due to several attributes: sensitivity, selectivity, multi-analyte analysis, and structural information. Increasing applications of mass spectrometry as quantitative measurement to assist in the evaluation and validation of biomarker leads has recently been reported [18]. A key to the accurate identification of these biomarkers is the ability to correctly distinguish the diseased mass spectra from the control by a classification scheme.

To demonstrate the performance of the proposed kriging-based possibilistic entropy on the classification of biosignals, we used a public MS-based ovarian cancer dataset, the ovarian high-resolution SELDI-TOF, to carry out the entropy analysis. The dataset was obtained from the FDA-NCI Clinical Proteomics Program Databank [19]. The ovarian cancer data consist of 100 control samples and 170 cancer samples. The length of each sample is 15,154 m/z values. We examined the complexity of this type of MS data and applied the possibilistic entropy PossEnP to classifying cancer and control samples and compared the performance of PossEnP with other methods. The classification of the diseased and control subspectra was carried out using the following rule: Assign subspectra i to class c_k (cancer or control) $\Leftrightarrow d_i = \min_j(d_{ij}), j \in c_k \forall k, i \neq j$, where $d_{ij} = (\mathbf{a}_i^T \mathbf{b}_i)/(\mathbf{a}_j^T \mathbf{b}_i)$; in which $\mathbf{a}_i$ is defined in Eq. (11) which is the kriging prediction vector of subspectra i, $\mathbf{b}_i$ is defined in Eq. (11) associated with i , and $\mathbf{a}_j$ is the kriging prediction vector of subspectra j. The value used for p is 6. The k-fold cross validations were then applied to measure the performance of the classification task.

Table 1. k-fold cross validation results for ovarian cancer data

k	μ_{cl}	σ_{cl}	μ_{cr}	σ_{cr}
SVM				
2	0.8930	0.0267	0.9492	0.0270
6	0.9094	0.0262	0.9760	0.0149
10	0.9096	0.0269	0.9801	0.0127
BK				
2	0.9320	0.0179	0.9721	0.0150
6	0.9496	0.0161	0.9858	0.0142
10	0.9560	0.0157	0.9902	0.0139
GeoEn				
2	0.9541	0.0145	0.9780	0.0144
6	0.9612	0.0140	0.9871	0.0138
10	0.9825	0.0138	0.9931	0.0135
PossEnP				
2	0.9702	0.0100	0.9870	0.0121
6	0.9789	0.0115	0.9901	0.0114
10	0.9914	0.0127	0.9975	0.0119

The SVM (support vector mahine) approach extracted the wavelet coefficients of the MS data as the features for cancer classification. The SVM-based classifier was reported to outperform several other classification algorithms including voted perceptron, discriminant analysis, decision tree analysis, naive Bayes, bagging and boosting classification trees, and random forest. The BK (block kriging) method applied a kriging scheme to estimate the error variance of a block of protein peaks of the MS data and used an error matching scheme for classification [20]. GeoEn [21] was applied to compute the prediction coefficients, and incorporated them to calculate the similarity measure for signal classification. The validation of the classification of the ovarian cancer was designed with similar strategies to those carried out in [22] using the k-fold cross validation, where k = 2, 6, and 10. For each k-fold validation, the averages were computed by the results of 1000 independent random experiments.

Using the same parameter setting described in [21], the performance of PossEnP was compared with previous results using the SVM, BK, and GeoEn. The control mean (μ_{cl}), cancer mean (μ_{cr}), control standard deviation (σ_{cl}), and cancer standard deviation (σ_{cr}) of the four methods are shown in Table 1. The results for sensitivity are higher than those for specificity found in all three methods. The sensitivity measures the percentage of actual positives (diseased) which are correctly identified as such; and the specificity measures the percentage of negatives (non-diseased) which are correctly identified. BK, GeoEn, and PossEnP provide results better than the SVM method. GeoEn outperformed BK in all k folds. Results obtained from PossEnP are found to be the best in all the k-fold validation outperforming SVM, BK, and GeoEn; including the means and standard deviations for control and cancer.

4. CONCLUSION

Although there are many nonlinear dynamical methods widely applied to physiological and biological signal analysis, including the approximate entropy family, for studying the complexity or predictability of time-series data; the proposed possibilistic entropy is the entropy-based approach that takes into account the usefulness of spatial information inherently existing in many types of time-series signals such as the mass spectrometry data in this study.

The proposed PossEnP can be useful for classification of complex biosignals in general, and further developed to be used as a computational tool for validation of potential protein biomarkers embedded in mass spectrometry data in particular.

5. REFERENCES

[1] C. Shannon, W. Weaver, *The Mathematical Theory of Information*. University of Illinois Press, 1949.

[2] A. De Luca, S. Termini, A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory, Information and Control 20 (1972) 301-312.

[3] A.N. Kolmogorov, Entropy per unit time as a metric invariant of automorphism, Doklady of Russian Academy of Sciences 124 (1959) 754-755.

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

[5] D. Kaplan, L. Glass, *Understanding Nonlinear Dynamics*. New York, Springer, 1995.

[6] P. Grassberger, I. Procaccia, Estimation of the Kolmogorov entropy from a chaotic signal, Phys Rev A 28 (1983) 2591-2593.

[7] J.P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, Rev Modern Phys 57 (1985) 617-654.

[8] N. Kannathal, M.L. Choo, U.R. Acharya, P.K. Sadasivan, Entropies for detection of epilepsy in EEG, Comput Meth Programs Biomed 80 (2005) 187-194.

[9] V. Srinivasan, C. Eswaran, N. Sriraam, Approximate entropy-based epileptic EEG detection using artificial neural networks, IEEE Trans Information Technology in Biomedicine 11 (2007) 288-295.

[10] J. S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Amer. J. Physiol. Heart Circ. Physiol. 278 (2000) H2039-H2049.

[11] M.-Y. Lee, C.-S. Yang, Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images, Comput Meth Programs Biomed 100 (2010) 269-282.

[12] W. Chen, J. Zhuang, W. Yu, Z. Wang, Measuring complexity using FuzzyEn, ApEn, and SampEn, Medical Engineering & Physics 31 (2009) 61-68.

[13] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets Syst 100 (1999) 9-34.

[14] T.D. Pham, Possibilistic entropy: A new approach for nonlinear dynamical analysis of biosignals, 15th Int. Conf. Knowledge-Based & Intelligent Information & Engineering Systems, LNAI 6881(2011) 466-473.

[15] E.H. Isaaks, R.M. Srivastava, *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.

[16] N. Cressie, C.K. Wikle, *Statistics for Spatio-Temporal Data*. Wiley, New Jersey, 2011.

[17] L. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*. New Jersey, Prentice Hall, 1993.

[18] B.L. Ackermann, J.E. Hale, K.L. Duffin, The role of mass spectrometry in biomarker discovery and measurement, Curr Drug Metab. 7 (2006) 525-539.

[19] http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

[20] T.D. Pham et al., Classification of mass spectrometry based protein markers by kriging error matching, Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry, LNAI 5108 (2008) 82-94.

[21] T.D. Pham, GeoEntropy: A measure of complexity and similarity, Pattern Recognition 43 (2010) 887-896.

[22] J.S.Yu et al., Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data, Bioinformatics 21 (2005) 2200-2209.