33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Interpretation of Coarse-Graining of Lempel-Ziv Complexity Measure in ECG Signal Analysis

Shijie Zhou, Zichen Zhang, *Student Member, IEEE* and Jason Gu, *Senior Member, IEEE*

Abstract—**Lempel-Ziv (LZ) complexity measure has been applied to classify ventricular tachycardia (VT) and ventricular fibrillation (VF). The coarse-graining process plays a crucial role in the LZ complexity measure analysis, which directly affects the separating performance of VT and VF in ECG signal analysis. The question of different coarse-graining approaches interpretability in ECG signal analysis and their influence on the performance of ECG classification have not yet been previously addressed in the literature. In this paper, we present four coarse-graining process approaches, K-Means, Mean, Median and Mid-point. Our test shows that K-Means algorithm is superior to the other three approaches in VT and VF separation rate, Particularly, optimum performance is achieved at a 8-second window length.**

I. Introduction

LEMPEL and Ziv (LZ) complexity measure is a useful approach based on a coarse-graining measurement to estimate the randomness of finite symbolic sequences to process the information data [1-2]. In the past decade, the LZ complexity measure (which offers relevant mathematical definitions and detailed derivations) has been extensively applied in biomedical signal analysis to evaluate the complexity of physiologic signals in discrete-time [3]. For example, EEG complexity measure in the depth of anesthesia [4-5], DNA sequences analysis [6] and classification of ECG signal [7-9].

In ECG signal analysis, the detection and classification of heart irregular rhythm plays an important role in diagnosing and preventing cardiovascular disease (CVD). Specifically, there are two high-risk arrhythmias, ventricular tachycardia (VT) and ventricular fibrillation (VF) that can be studied to understand the pathological changes and biological mechanisms of deadly CVD. VF has been considered a chaotic state (a random and irregular process). VT is a periodic motion exhibiting a rapid heart rate, giving rise to a diminished cardiac output when it occurs. In non-linear signal processing, the LZ complexity measure can cope with a dynamical system entering a chaotic state by quantifying the rate of new pattern occurrences along given finite symbolic sequences. Therefore, according to the features of arrhythmias mentioned above, the LZ complexity measure has been adopted to classify the VT and VF, presented by previous research that the distributions of VT and VF could be exactly separated in their own database [7-9].

The LZ complexity measure includes several steps in separating the arrhythmia. Firstly, the original signal is extracted from an ECG device. Then the ECG signal of an appropriate window length should be converted to a finite symbolic sequence using coarse-graining technique. Lastly, the LZ complexity analysis counts the number of new patterns in the given finite symbolic sequence scanned from left to right. The coarse-graining process determines how much inherent information can be retained and will consequently impact the separation of VT and VF. Although the original signal loses a large amount of information in the process, the inherent features of heart dynamics are reserved, and robustness to noise is guaranteed. The question of different coarse-graining approaches interpretability in ECG signal analysis and their influence on the performance of ECG classification have not yet been previously addressed in the literature. We have utilized four methods (K-Means, Mean, Mid-point and Median) in coarse-graining process aiming at gaining a better understanding of their impact on the classification of ECG signal.

The paper is organized as follows. In Section 2, different methods in the coarse-graining process are described. In Section 3, we present the test database and discuss the simulation results. Conclusions are given in Section 4.

II. COARSE-GRAINING TECHNIQUE

Coarse-graining process has been used successfully to quantify the original signals in given finite symbolic sequence. With a selected window length, the finite original signal is transformed into a symbolic sequence corresponding to consecutive non-overlapping time intervals of sizeΔ. In this paper, four methods are applied in coarse-graining process and their performances are evaluated.

A. K-Means

Rapp and et al. [10] proposed several techniques to map a time series into a sequence of symbols, which is referred to as "partitioning". The procedure of transforming the 1-lead ECG signal into a binary sequence using the k-means algorithm can be described as follows.

1) Select the window length (WL).

Shijie Zhou is with the Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS B3J 2X4, Canada (e-mail: shijiezhou@dal.ca).

Zichen Zhang is with the Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS B3J 2X4, Canada (e-mail: zichen.zhang@dal.ca).

Jason Gu is with Electronic & Information Engineering College, Henan University of Science and Technology, Henan, China and also the School of Biomedical Engineering, and the Department of Electrical and Computer, Dalhousie University, Halifax, NS B3J 2X4, Canada (corresponding author to provide phone: 1-902-494-3163; e-mail: Jason.Gu@dal.ca).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

2) Iteration 0, Let $K = 2$, Two new centroids $z_1(1),\ z_2(1)$ can be produced by a quantity of center (assume $\varepsilon = 0.005$):
($x_m + \varepsilon * x_m$ and $x_m - \varepsilon * x_m$). , where x_m is the mean value of the data points $\{x_i |\ i = 1, 2, 3, \dots, n\}$ within the selected window length and $n = Sample\ Rate * WL$. It can be estimated as

$$x_m = (1/n) \sum_{i=1}^{n} x_i$$

3) Iteration 0, because the ECG signal is only 1 lead signal, we can calculate the distance between the centroid to each data point, For $i = 1, 2, 3, \dots, n$, the distances are.

$$D_1^i = \parallel x_i - \mathrm{z}_1(1) \parallel^2$$

$$D_2^i = \parallel x_i - \mathrm{z}_2(1) \parallel^2$$

4) Iteration 0, each data point based on the minimum distance is set 1, $i = 1, 2, 3, \dots, n$.

If $D_1^i < D_2^i$,
$x_i = 1$ is assigned to group 1,
Otherwise
$x_i = 0$ is assigned to group 2.

5) Iteration 1, Determine two new centroids: knowing the two groups, now the new centroid of each group can be obtained. Group 1, the center $z_1(2)$ is the average coordinate among all of members in the group 1. Group 2, the center $z_2(2)$ is the average coordinate among all of members in this group.

6) Iteration 1, the distance between the new centroid to each data point is calculated.

$$D_1^i = \parallel x_i - \mathrm{z}_1(2) \parallel^2$$

$$D_2^i = \parallel x_i - \mathrm{z}_2(2) \parallel^2$$

7) Iteration 1, repeat step 4.

8) Repeat the above procedure until $z_1(k+1) = z_2(k)$ for $k = 1, 2, \dots,$.

B. Mean

Zhu *etc.* [3] utilized the mean value of the data in the partitioning process, which is described as follows:

1) Select the window length (WL).

2) The mean value of the data points $\{\ x_i |\ i = 1, 2, 3, \dots, n\ \}$ within the selected window length can be estimated as

$$x_m = (1/n) \sum_{i=1}^{n} x_i$$

where $n = Sample\ Rate * WL$.

3) The average x_m is subtracted from every data point $(x_i - x_m)$, and then obtain the positive peak value V_p and negative peak value V_n.

4) Count the P_c and N_c, where P_c represents the number of data x_i ($0 < x_i < 10\%V_p$), and N_c represents the number of data x_i ($10\%V_n < x_i < 0$).

5) Select the threshold T_d , when $P_c + N_c < 40\%n$, the $T_d = 0$, if $P_c < N_c$, the $T_d = 20\%V_p$, else the $T_d = 20\%V_n$.

6) Finally, compared with the threshold T_d, the data are turned into a binary sequence $S_1, S_2, S_3, \dots, S_r$. If $x_i < T_d, S_i = 0$, otherwise the $S_i = 1$.

These given parameters "10%", "40%" and "20%" were derived by an empirical study of ECG data [3].

C. Median

The median value is determined by the entire data set, the finite sequence should be sorted and then select the one in the middle.

1) Select the length of data segment, i.e., window length (WL), where $n = Sample\ Rate * WL$.

2) Sort the data in ascending order, and then select the one in the middle.

3) To subtract each data point with the median value ($x_i - x_{median}$), then find the positive peak value V_p and negative peak value V_n.

Then steps 4-6 are the exactly the same as explained in steps 4-6 of Part B.

D. Mid-Point

The Mid-point is used to transform the each data point into the binary sequence as follows, which is determined by only two data.

1) Select the length of data segment.

2) The mid-point value of the data points $\{\ x_i |\ i = 1, 2, 3, \dots, n\}$ within the selected window length. The smallest value $x_{min}$and the largest value x_{max},

$$x_{m-point} = \frac{x_{min} + x_{max}}{2}$$

3) Obtain the positive peak value $V_p$and negative peak value V_n by the average $x_{m-point}$ is subtracted by the every data point $(x_i - x_{m-point})$.

Then steps 4-6 are the exactly the same as explained in steps 4-6 of Part B.

III. SIMULATION AND RESULTS

A. Band-pass Filter

The band-pass filter preceding the coarse-graining process is used to reduce the influence of baseline wander, different types of noise, and so on. Integer filter is applied in our study to minimize the computational cost [11]. The pass-band frequency is about 0.6Hz - 22Hz, to maximize the energy of the signal that is of interest (SR, VT and VF). The cut-off frequency of the high-pass filter is 0.6 Hz, corresponding to a delay of 127.5 samples. And the cut-off frequency 22 Hz for the low-pass filter corresponds to a delay of 2 samples. In order to uniform the sample rate in the study, the sample rate is changed to 200Hz.

B. *LZ complexity analysis*

After passing through the filter, the signal must firstly be transformed into a finite binary symbol sequence known as a coarse-graining process, and then the complexity measure can be gained by counting the number of the distinct patterns in the given binary sequence with the discrete time series. In other words, the LZ complexity analysis is utilized to calculate the new pattern generation rate in the given finite sequence. In the context of ECG signal analysis, the signal $x(n)$ is converted into a binary sequence which will then be scanned from left to right. Throughout the whole sequence, the complexity counter $c(n)$ is increased by one when a new subsequence of consecutive binary sequence is encountered in the scanning process. Finally, the $C(n)$ denoted the normalized output of LZ complexity analysis instead of $c(n)$ is used to describe the results. The LZ complexity analysis is described in [2, 3]

C. *Simulation Data*

A set of ECG records obtained from the MIT-BIH database [13]. In [14], in order to obtain $C(n)$ results which are independent of n, the finite string needs at least $n > 1000$. Hence, the window length was set from 5 seconds beginning. We selected 6 different window lengths to analyze the impact of window length on the coarse-graining process. The VT has a wide range, which includes both monomorphic and polymorphic types. Therefore, it is very difficult to distinguish VT from VF; causing the main error in previous studies. For each selected window length in our testing, 196 monomorphic VT segments and 114 VF segments obtained from the Malignant Arrhythmia subset of MIT-BIH are used for both development and evaluation stages.

1) *Development Stage*

There are 98VT segments and 57 VF segments selected from the database for development stage, and the sampling frequency was set to 200 Hz. Using different coarse-graining process approaches, we used these data to obtain the threshold for the evaluation stage.

2) *Evaluation Stage*

For evaluation, we selected 100 sinus rhythm (SR) segments, 98 VT segments and 57 VF segments from the *MIT-BIH Database* at each different window length. The $C(n)$ value obtained from the development stage is used to evaluate the sensitivity of SR, VT and VF based on different coarse-graining process approaches. It is then used to interpret the impact of the coarse-graining process in ECG signal analysis.

D. *Results*

Four coarse-graining approaches are studied here: k-Means, means, median and mid-point.

By examining the probability density function (PDF) which is shown in Fig. 2, a threshold for distinguishing between VT and VF is found. The signal is considered to be VT if $\mathrm{C(n)}$ is less than $\mathrm{C(n)_T}$, otherwise, it is classified as VF.

Figure 3 illustrates the test results. k-means algorithm outperforms the other three methods in terms of classification of VT and VF. This implies that the symbolic sequence constructed by k-means partitioning is a better representation of the original signal. In addition, the effect of different window lengths was evaluated. The best sensitivity was obtained at the 8-second window length. Note that, in development stage where we chose the threshold, mean-based approach gives a clear trend of distribution of VT and VF, as can been seen in figure 2(b). However, it yields a much lower sensitivity in the evaluation stage.

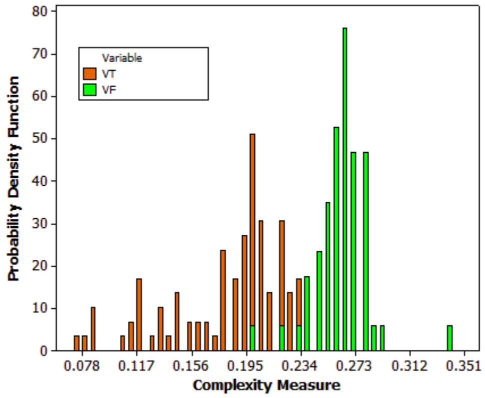


Fig. 2 (a): K-Means with 8 sec window length

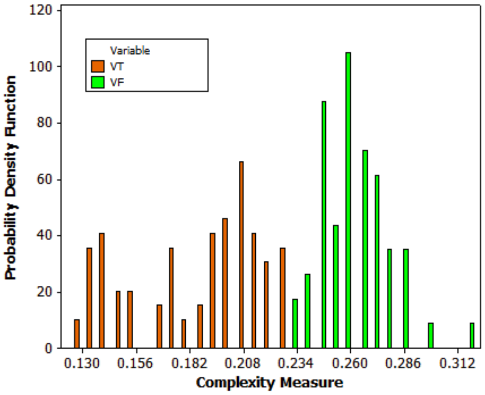


Fig. 2 (b): Mean with 8 sec window length

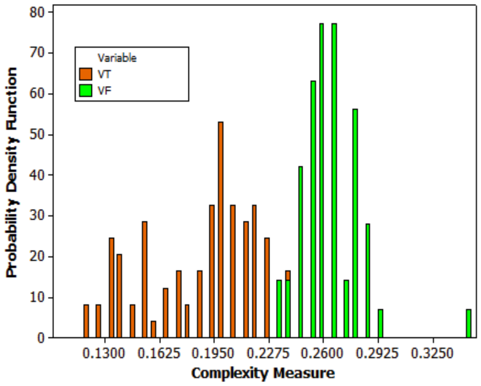


Fig. 2 (c): Median with 8 sec window length

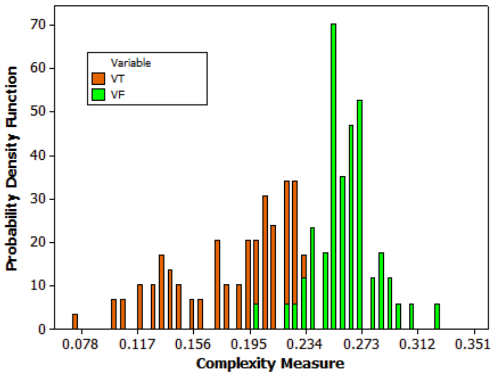


Fig. 2 (d): Mid-point with 8 sec window length

IV. Conclusion

In this paper, we studied how different coarse-graining methods (k-means, mean, median, mid-point) influence the result of LZ complexity analysis. Time is a crucial issue for the detection and classification of VT/VF. A sensitive algorithm to detect and classify the VT/VF will efficiently improve the survival probability of the patients. The LZ complexity measure shows itself possessing good advantages in time domain analysis. LZ is particularly useful in describing the complexity of random processes. We used the information theory to analyze the chaotic ECG signal. Coarse-graining process can be used to quantify the signal. The central idea is to transform the original data into a finite binary sequence. Different coarse-graining methods are used to fully understand how they affect the ECG classification. The binary time series sequence obtained by using k-means algorithm is found to be a better representation of the original signal, and then a suitable threshold of LZ complexity can be easily obtained, with which VT and VF can be efficiently classified. Our results not only show that a proper coarse-graining technique is essential in the success of LZ complexity analysis in ECG signal classification, but also suggest k-mean algorithm as a better method in coarse-graining compared to mean, median and mid-point.

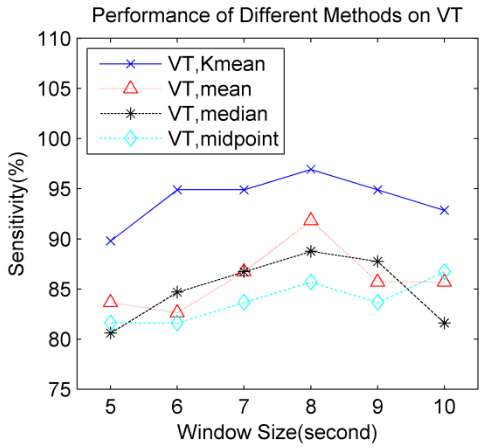
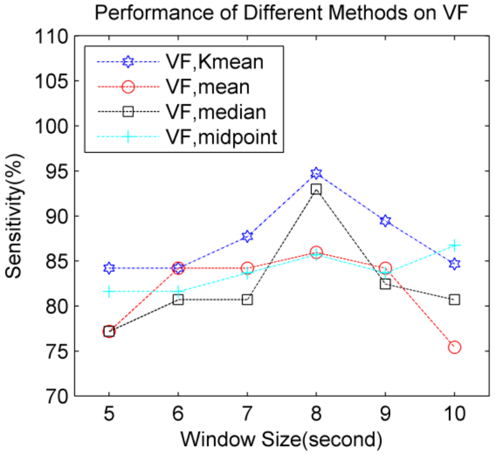


Fig. 3(a) Performance of different coarse-graining process approaches at different window lengths for VT



(b)

Fig. 3(b) Performance of different coarse-graining process approaches at different window lengths for VF

$$* \ \text{Sensitivity}\% = \text{TP}/(\text{TP}+\text{FN}) \text{ , where TP} = \text{true positive, FN} = \text{false negative [12]}$$

References

[1] A. Lempel and J. Ziv, "On the complexity of finite sequences," IEEE Trans. Inf. Theory, vol. IT-22, no. 1, pp. 75–81, 1976.

[2] J. Ziv and N.Merhav, "Estimating the number of states of a finite-state source," IEEE Trans. Inf. Theory, vol. 38, no. 1, pp.61–65, 1992.

[3] H. X. Zhang, Y. S. Zhu, and Z. M. Wang, "Complexity measure and complexity rate information based detection of ventricular tachycardia and fibrillation," Med. Biol. Eng. Comput., vol. 38, no. 5, pp. 553–557, Sep. 2000.

[4] X. S. Zhang and R. J. Roy, "Derived fuzzy knowledge model for estimating the depth of anesthesia," IEEE Trans. Biomed. Eng., vol. 48, no. 3, pp. 312–323, Mar. 2001

[5] X. S. Zhang, Y. S. Zhu, and X. J. Zhang, "New approach to studies on ECG dynamics: extraction and analyses of QRS complex irregularity time series,"Med. Biol. Eng. Comput., vol. 35, no. 5, pp. 467–473, Sep.1997.

[6] X. S. Zhang, R. J. Roy, and E. W. Jensen, "EEG complexity as a measure of depth of anesthesia for patients," IEEE Trans. Biomed. Eng., vol. 48, no. 12, pp. 1424–1433, Dec. 2001.

[7] X. S. Zhang, Y. S. Zhu, N. V. Thakor, and Z. Z. Wang, "Detecting ventricular tachycardia and fibrillation by complexity measure," IEEE Trans. Biomed. Eng., vol. 46, no. 5, pp. 548–555, May 1999.

[8] Ayesta U., Serrano L., Romero I.," Complexity Measure revisited: A new algorithm for classifying cardiac arrhythmias", 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2001).

[9] Yinlin Xu, Qianli D.Y. Ma and Daniel T. Schmitt, etc., "Effects of coarse-graining on the scaling behavior of long-range correlated and anti-correlated signals", physics. data-an, Feb 19, 2010

[10] Linde, Y., Buzo, A. & Gray, R. M., "An algorithm for vector quantizer design," IEEE Trans. 1980, Commun.28, pp.84-95.

[11] T. Luczak andW. Szpankowski, "A suboptimal lossy data compression based on approximate pattern matching," IEEE Trans. Inf. Theory, vol.43, no. 5, pp. 1439–1451, Sep. 1997.

[12] N.V. Thakor, "From Holter monitors to automatic defibrillators: Developments in ambulatory arrhythmia monitoring", IEEE Trans. Biomed Eng., Vol. BME-31, pp 770-779, 1984.

[13] MIT-BIH database, Available website [June 18, 2011] http://physionet.org/physiobank/database/

[14] F. Kaspar and H. G. Schuster, "Easily calculable measure for the complexity of spatiotemporal patterns," Phys. Rev. A., vol. 36, pp.842–848, 1987.