Belief Kullback–Leibler Divergence-based Dynamical Complexity Analysis for Biological Systems

Lang Zhang and Fuyuan Xiao*

Chongqing University, Chongqing, China

The 10th International Conference on Information Systems and Computing Technology (ISCT)

Information Processing and Intelligent Systems Lab

Outline

01 **Background & Motivation**

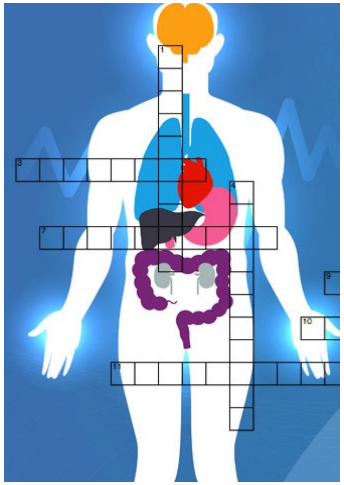
02 **Preliminaries**

03 **The Proposed Method**

04 **Simulation Experiments**

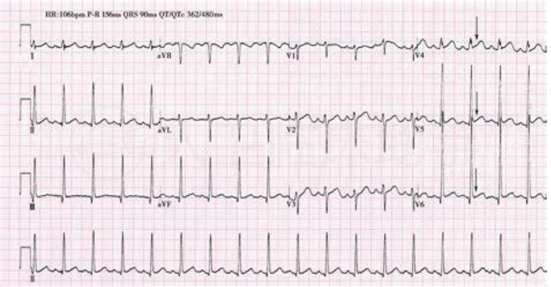
05 **Conclusion**

Background



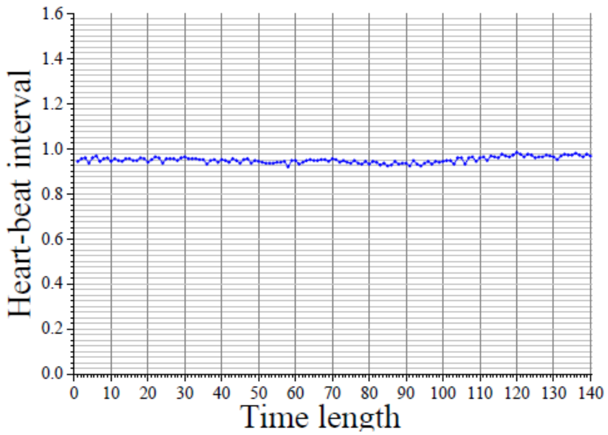
Biological systems

EEG signals



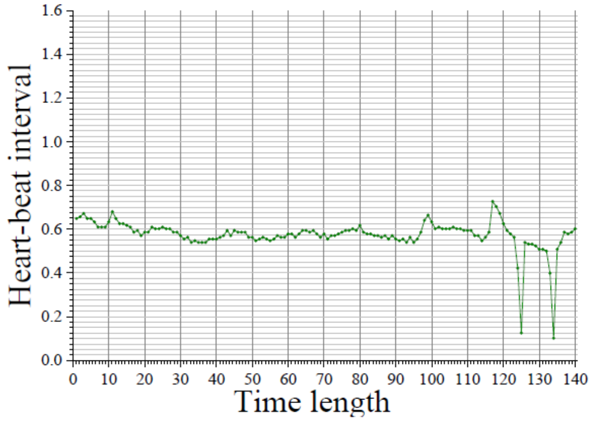
Cardiac inter-beat signals

Background



Pathological subjects

More information →



Healthy subjects

- **Recent works for uncertain information management:**

Deng entropy

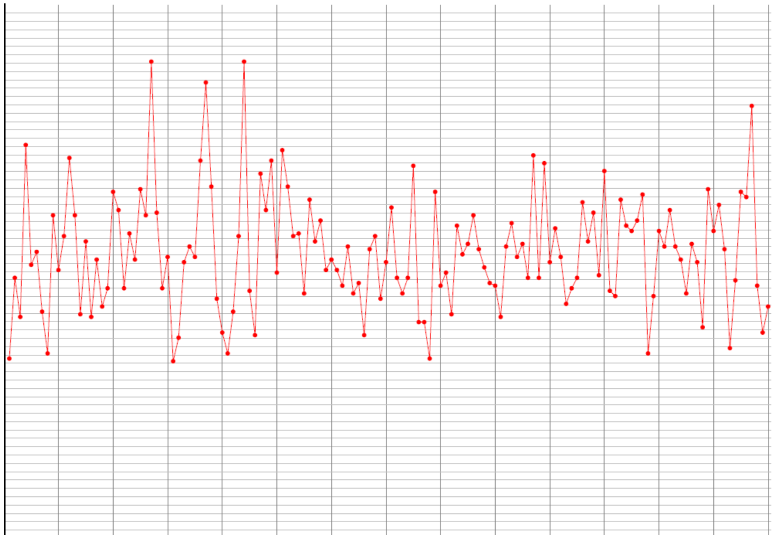
Z-network

Information quality

Belief entropy

TDBF

Motivation



Sophisticated structure

Feature extraction based on Dempster-Shafer evidence theory

BPAs

Process the time series data points whether they are on the boundaries of time slice

Preliminaries

- **Dempster-Shafer evidence theory**

Definition 1 (Framework of discernment).

Let Θ be a set that consists of r mutually exclusive and collectively exhaustive events,

$$\Theta = \{e_1, e_2, \dots, e_i, \dots, e_r\}$$

which indicates the framework of discernment. The power set 2^Θ is used to describe uncertainty which can be defined as follows:

$$2^\Theta = \{\emptyset, \{e_1\}, \dots, \{e_r\}, \{e_1, e_2\}, \dots, \{e_1, e_2, \dots, e_h\}, \dots, \Theta\},$$

where $\emptyset$ indicates the empty set.

Definition 2 (Mass function).

Based on the frame of discernment Θ, m as a mass function, also known as BPA, is a mapping from 2^Θ to [0,1] which is defined as:

$$m: 2^\Theta \to [0,1].$$

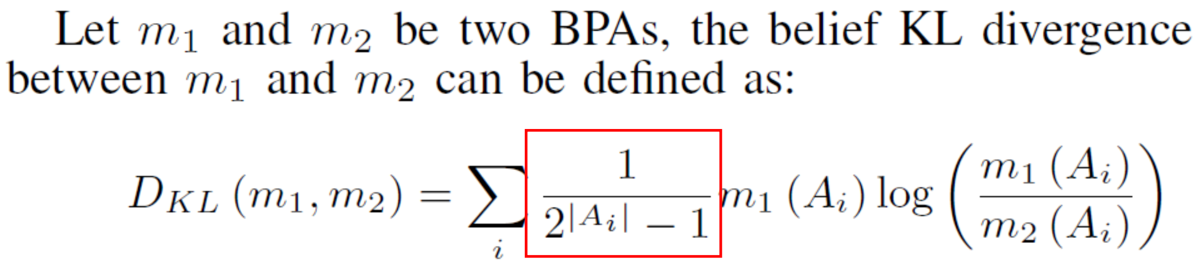
Because events must arise from propositions in the framework of discernment and empty set is not the cause of the events, it abides the rule of

$$\sum_{E \in 2^\Theta} m(E) = 1 \text{ and } m(\emptyset) = 0.$$

If $m(E) > 0$ E is a focal element.

Preliminaries

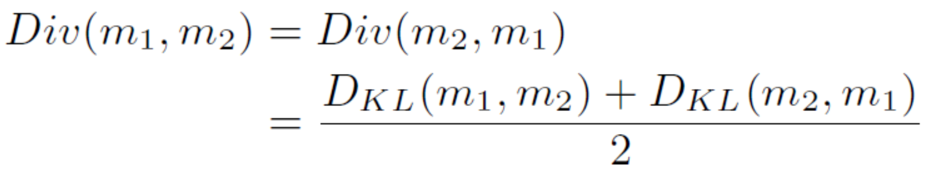
- **Belief KL divergence measure**



Let m_1 and m_2 be two BPAs, the belief KL divergence between m_1 and m_2 can be defined as:

$$D_{KL}\left(m_1, m_2\right) = \sum_i \boxed{\frac{1}{2^{|A_i|} - 1}} m_1\left(A_i\right) \log\left(\frac{m_1\left(A_i\right)}{m_2\left(A_i\right)}\right)$$

- **Make it be symmetric**



$$\begin{aligned} Div(m_1, m_2) &= Div(m_2, m_1) \\ &= \frac{D_{KL}(m_1, m_2) + D_{KL}(m_2, m_1)}{2} \end{aligned}$$

The Proposed Method （BKLDC）

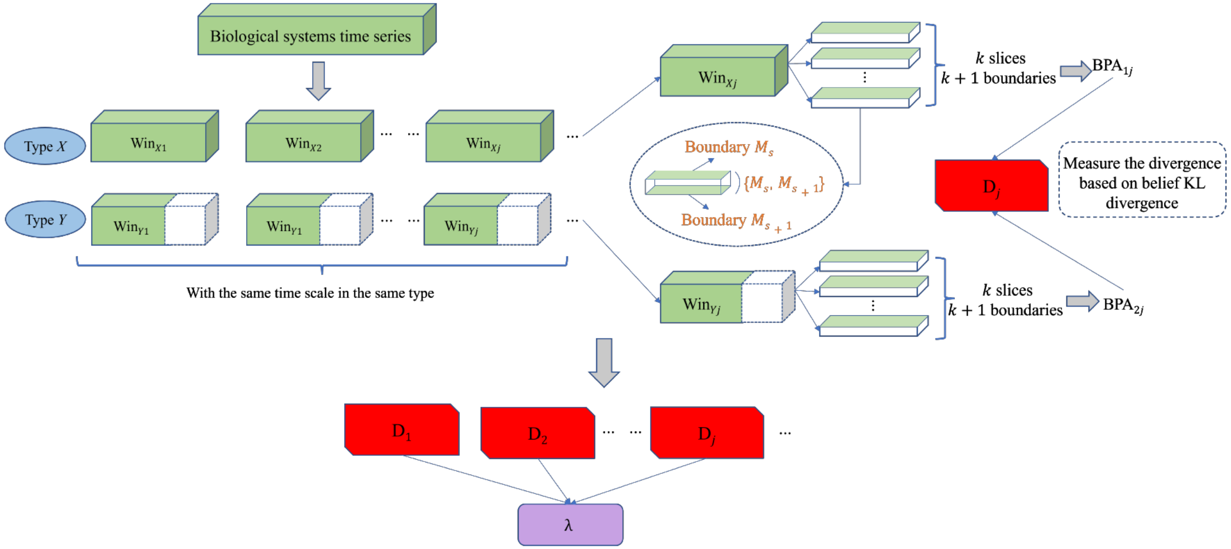


Figure 1. Flowchart of the BKLDC algorithm for biological systems.

- **Two lists of consecutive non-overlapping time windows**

$$w_{Xj}^{(\eta)} = \left\{t_{(j-1)\eta+1}, \ldots, t_{(j-1)\eta+\eta}\right\}$$

$$w_{Yj}^{(\eta)} = \left\{t_{(j-1)\eta+1}, \ldots, t_{(j-1)\eta+\delta}\right\}$$

- **Focal element of BPA**

$$m_{ij}: \begin{cases} m_{ij}^{(\eta)}(\{M_s\}) = \dfrac{q_{ij}}{|w_{ij}|}, & \text{if } \gamma \text{ falls on the boundary } s, \\ m_{ij}^{(\eta)}(\{M_s, M_{s+1}\}) = \dfrac{p_{ij}}{|w_{ij}|}, & \text{otherwise,} \end{cases}$$

- ***D$_j$* in each corresponding window**

$$\begin{aligned} D_j^{(\eta)} &= Div(m_{Xj}, m_{Yj}) \\ &= \frac{1}{2} \cdot \sum_{s=1}^{k} \frac{m_{Xj}(\{M_s, M_{s+1}\})}{2^{|\{M_s, M_{s+1}\}|} - 1} \log \left(\frac{m_{Xj}(\{M_s, M_{s+1}\})}{m_{Yj}(\{M_s, M_{s+1}\})} \right) \\ &+ \frac{1}{2} \cdot \sum_{s=1}^{k} \frac{m_{Yj}(\{M_s, M_{s+1}\})}{2^{|\{M_s, M_{s+1}\}|} - 1} \log \left(\frac{m_{Yj}(\{M_s, M_{s+1}\})}{m_{Xj}(\{M_s, M_{s+1}\})} \right) \\ &= \frac{1}{2} \cdot \sum_{s=1}^{k} \left(\frac{p_{Xj}}{3 \cdot |w_{Xj}|} - \frac{p_{Yj}}{3 \cdot |w_{Yj}|} \right) \log \frac{p_{Xj} \cdot |w_{Yj}|}{p_{Yj} \cdot |w_{Xj}|}. \end{aligned}$$

- **Property 1.** When all the data fall on the boundaries

$$
\begin{aligned}
D_j^{(\eta)} &= Div(m_{Xj}, m_{Yj}) \\
&= \frac{1}{2} \cdot \sum_{s=1}^{k} \left(\frac{q_{Xj}}{|w_{Xj}|} - \frac{q_{Yj}}{|w_{Yj}|} \right) \log \frac{q_{Xj} \cdot |w_{Yj}|}{q_{Yj} \cdot |w_{Xj}|}.
\end{aligned}
$$

- **Property 2.** $D_j^{(\eta)}$ is symmetric as:

$$D_j^{(\eta)} = Div(m_{Xj}, m_{Yj}) = Div(m_{Yj}, m_{Xj})$$

- **Property 3.** When $m_{Xj} = m_{Yj}$, the $D_j^{(\eta)}$ is regarded as:

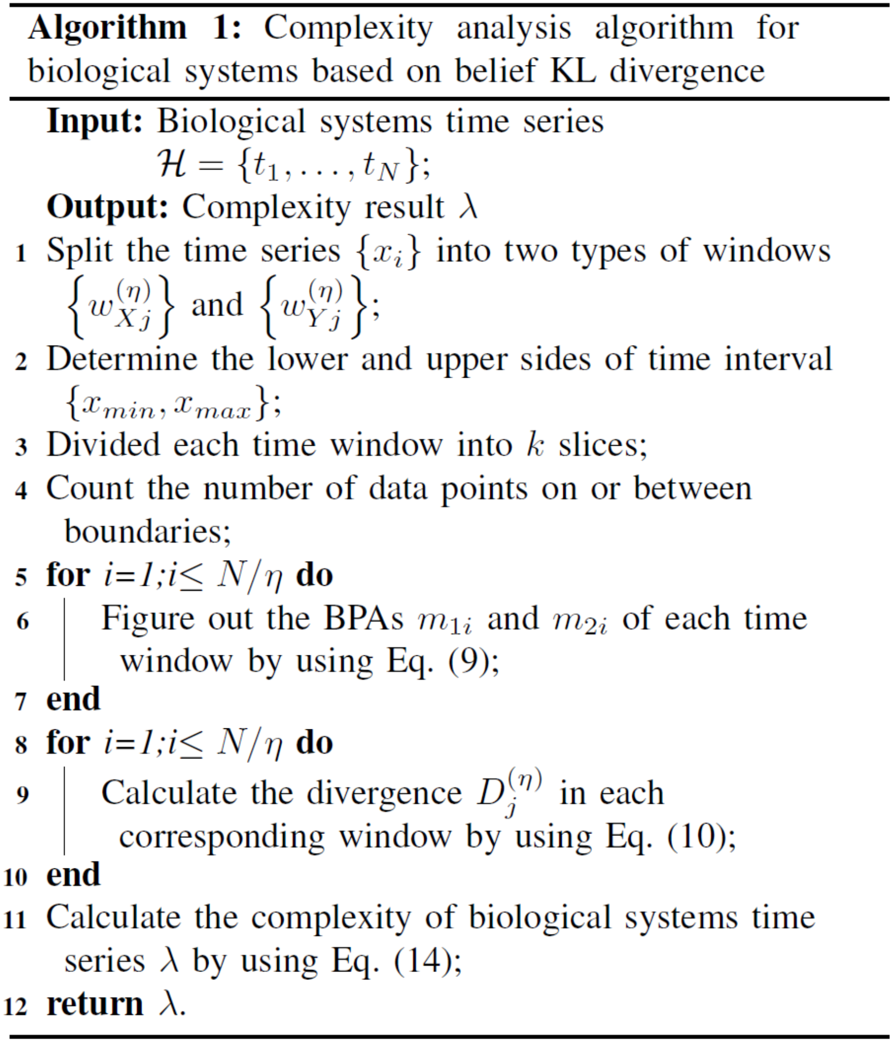
$$D_j^{(\eta)} = 0.$$

- **The average divergence represents the complexity of a biological system time series λ**

$$\lambda = \frac{\sum_{i=1}^{N/\eta} D_j^{(\eta)}}{N/\eta}.$$

The Proposed Method

The pseudocode of dynamical complexity analysis algorithm for biological systems based on KL divergence is shown in Algorithm 1.



Algorithm 1: Complexity analysis algorithm for biological systems based on belief KL divergence

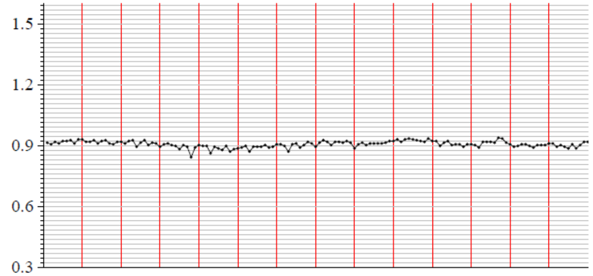
```
Input: Biological systems time series
       $\mathcal{H} = \{t_1, \ldots, t_N\}$;
Output: Complexity result $\lambda$
Split the time series $\{x_i\}$ into two types of windows
  $\left\{w_{Xj}^{(\eta)}\right\}$ and $\left\{w_{Yj}^{(\eta)}\right\}$;
Determine the lower and upper sides of time interval
  $\{x_{min}, x_{max}\}$;
Divided each time window into $k$ slices;
Count the number of data points on or between
  boundaries;
for $i$=1;$i \leq N/\eta$ do
    Figure out the BPAs $m_{1i}$ and $m_{2i}$ of each time
      window by using Eq. (9);
end
for $i$=1;$i \leq N/\eta$ do
    Calculate the divergence $D_j^{(\eta)}$ in each
      corresponding window by using Eq. (10);
end
Calculate the complexity of biological systems time
  series $\lambda$ by using Eq. (14);
return $\lambda$.
```

Feature extraction

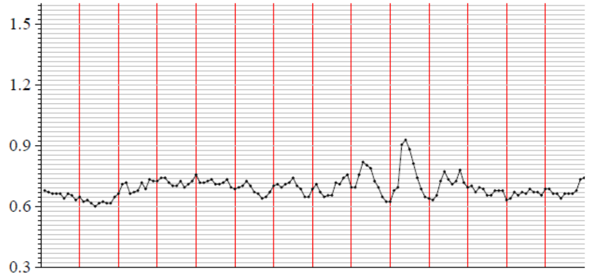
Complexity measurement

Simulation Experiments

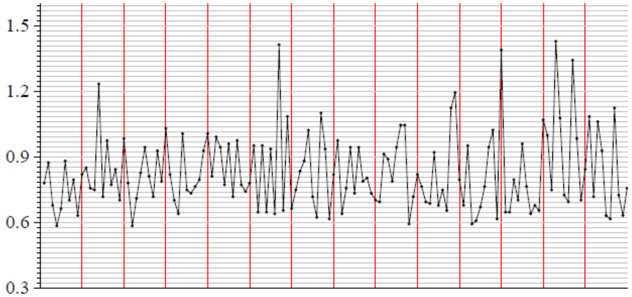
Cardiac inter-beat time series



Original time series of CHF



Original time series of Healthy



Original time series of AF

In this study, cardiac inter-beat interval time series is applied to demonstrate the feasibility of BKLDC algorithm for biological systems complexity analysis. The data is selected from the databases on PhysioNet as follows:

- BIDMC Congestive Heart Failure Database (CHF);
- MIT-BIH Normal Sinus Rhythm Database (Healthy);
- Long Term AF Database (AF).

Simulation Experiments

Data processing

- Each subject is truncated into 5 sets inter-beat interval time series by utilizing the first 500 data points from 10,000 data points.

- Hence, there are 240 sets inter-beat interval time series. Specifically, 75, 90 and 75 records are from CHF Healthy and AF, respectively.

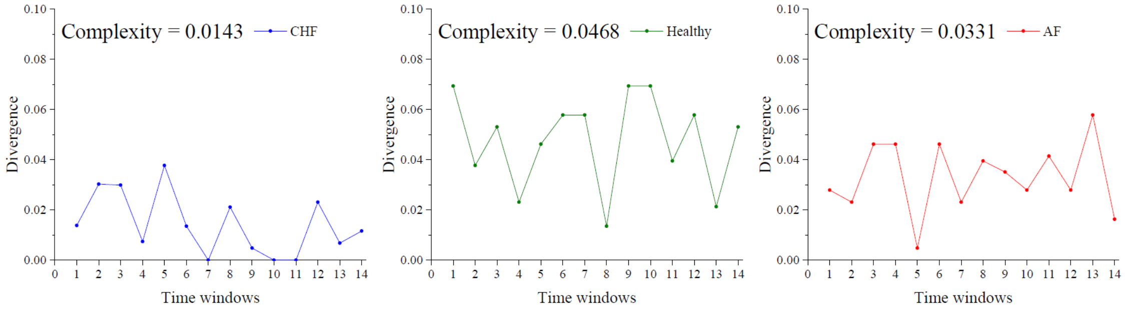
- The data points $\{x_i\}$ are ranked and split into 1000 segments. To release the influence of noise and detection error, the $1st$ and $999th$ 1000-quantiles of the ranked segments are regarded as x_{min}= 0.3 and x_{max} = 1.6

Simulation Experiments

- Table 1 shows the 14 divergence values for each time window of data sets, respectively.

Subject	Win_1	Win_2	Win_3	Win_4	Win_5	Win_6	Win_7	Win_8	Win_9	Win_{10}	Win_{11}	Win_{12}	Win_{13}	Win_{14}
CHF	0.0138	0.0303	0.0299	0.0074	0.0377	0.0135	0.0000	0.0211	0.0048	0.0000	0.0000	0.0231	0.0068	0.0116
Healthy	0.0693	0.0377	0.0530	0.0231	0.0462	0.0578	0.0578	0.0135	0.0693	0.0693	0.0395	0.0578	0.0213	0.0530
AF	0.0279	0.0231	0.0462	0.0462	0.0048	0.0462	0.0231	0.0395	0.0351	0.0279	0.0414	0.0279	0.0578	0.0163

- Fig. 3 shows the three original time series and divergence series, respectively.



(a) CHF subject

(b) Healthy subject

(c) AF subject

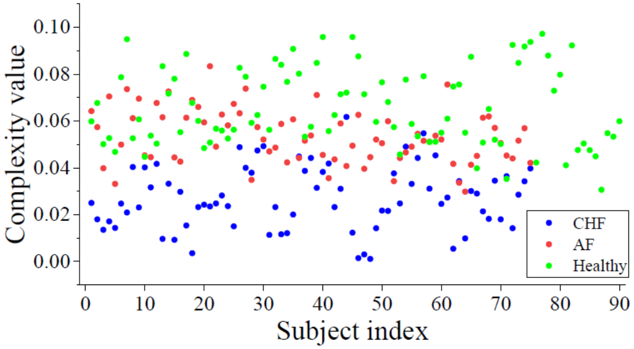


Figure 4. The complexity value in 240 sets cardiac inter-beat interval time series.

Simulation Experiments

- Sensitivity of pathological subjects and specificity of healthy subjects are defined as follows：

$$\text{Specificity:}V_{sp} = \frac{T_H}{T_H + F_H},$$

$$\text{Sensitivity:}V_{se} = \frac{T_P}{T_P + F_P},$$

$$\text{Accuracy:}V_{ac} = \frac{T_H + T_P}{T_H + F_H + T_P + F_P},$$

where T_H and F_H represent the amount of healthy subjects that classified correctly and falsely, respectively. Besides, T_P and F_P represent the amount of pathology subjects that classified correctly and falsely.

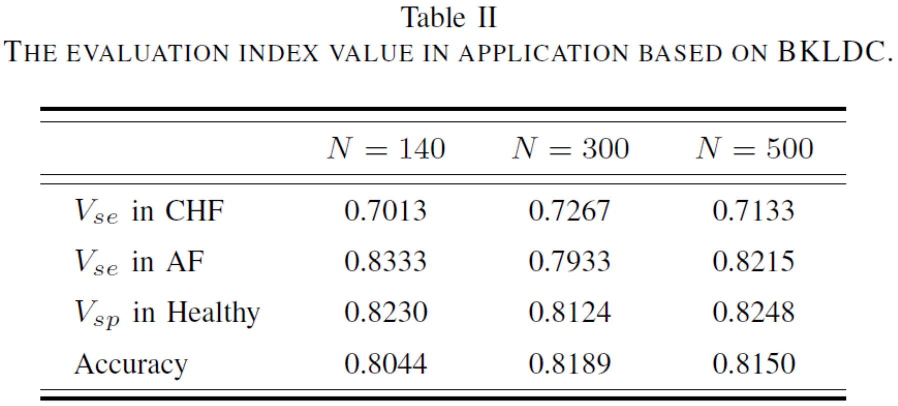
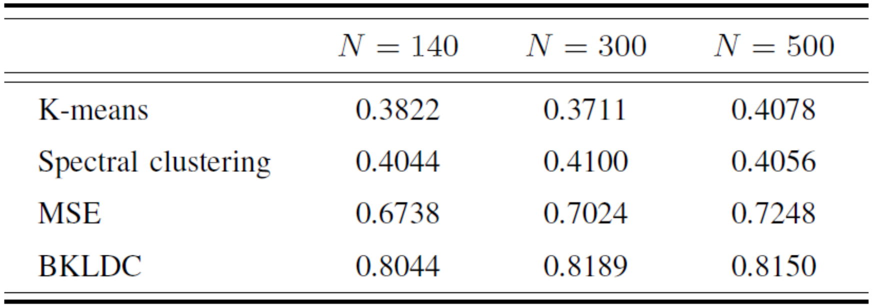


Table II
THE EVALUATION INDEX VALUE IN APPLICATION BASED ON BKLDC.

	$N = 140$	$N = 300$	$N = 500$
V_{se} in CHF	0.7013	0.7267	0.7133
V_{se} in AF	0.8333	0.7933	0.8215
V_{sp} in Healthy	0.8230	0.8124	0.8248
Accuracy	0.8044	0.8189	0.8150

Simulation Experiments

- The patter classification accuracy based on different methods.



	$N = 140$	$N = 300$	$N = 500$
K-means	0.3822	0.3711	0.4078
Spectral clustering	0.4044	0.4100	0.4056
MSE	0.6738	0.7024	0.7248
BKLDC	0.8044	0.8189	0.8150

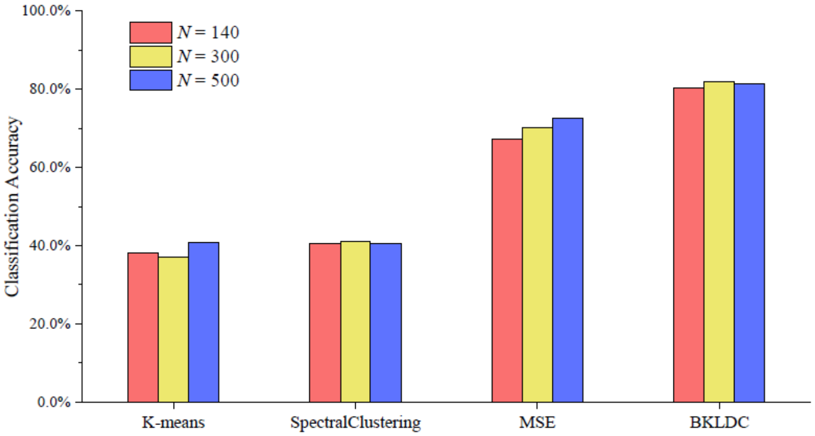


Figure 5. The pattern classification accuracy in cardiac inter-beat interval time series.

Conclusion

Contribution:

- Biological systems time series data is converted into mass function by using the D-S evidence theory, where feature of data can be extracted.
- The proposed BKLDC algorithm proposes an effective way to figure out the complexity of time series data in biological systems by generating BPAs and measure the average divergence of them.
- An application for pathological states analysis in cardiac inter-beat interval time series is carried out to illustrate the effectiveness of BKLDC algorithm.

Future work:

- The time complexity of the BKLDC algorithm for biological systems should be addressed to adapt to real-time data flexibly.

THANK YOU VERY MUCH!

Any questions and comments are welcome!

Email address: zhanglang_cqu@163.com