Discovering Optimal Variable-length Time Series Motifs in Large-Scale Wearable Recordings of Human Bio-behavioral Signals

Tiantian Feng, Shrikanth Narayanan

1Signal Analysis and Interpretation Lab (SAIL), University of Southern California

Summary

Background:
- Motifs are repetitive similar patterns that frequently appear in time-series.
- Motifs existing in wearable sensor signals can help to understand bio-behavioral patterns such as sleep pattern, commute behavior.

Challenges:
- How to identify the varying motif length?
- How to identify the optimal set of motifs?

Motifs are repetitive similar patterns that frequently appear in time-series.

Method

Data Aggregation
- Pre-processing
- Data Filtering
- HIME Motif Algorithm
- Top-K Motif Extraction

Figure 3: Optimal variable-length time-series motif learning pipeline.

Pre-processing
- Data aggregation: The time-series is aggregated every 3 minute.
- Data filtering: Savitzky-Golay filter.
- HIME (Hierarchical based Motif Enumeration): Variable-length motifs are detected using in a single SAX word $w = w_1w_2...w_l$.

Figure 4: How should we choose the top-K motifs?

Top-K Motif Learning
- Matrix $U$: PAA representation of $J$ subsequences ($U \in \mathbb{R}^{|p-q|}$), where each row in $U$ represents a PAA vector of one subsequence of length $L$.
- Define total frequency given $M$ motifs and distance threshold $D$:

$$F(M) = \sum_{k=1}^{K} \sum_{j=1}^{J} F_{kj},$$

where $F_{kj} = \begin{cases} 1, & \text{if } \sum_{l=1}^{L}(M_{kj} - U_{jl})^2 < D \\ 0, & \text{otherwise} \end{cases}$

- We want function $F(M)$ is maximized while $M$ motifs are different from each other by $2 \times D$:

$$M^* = \arg\max_{M} F(M), \text{subject to } \sum_{l=1}^{L}(M_{kj} - U_{jl})^2 > 2D$$

- We approximate $F(M)$ using a Gaussian kernel to handle zero derivative of $F(M)$ and discontinuity at the point where $\sum_{l=1}^{L}(M_{kj} - U_{jl})^2 = 0$.
- Apply gradient ascent solution to solve this optimization problem.

Figure 2: Example of Generating SAX word.

Figure 6: 4 subsequences of a motif that occurred repeatedly during the end of sleep

TILES Dataset

- TILES, Tracking Individual pErformance with Sensors, is a comprehensive human-subject experiments conducted in early 2018 to examine how the physiological, environmental, and behavioral variables impact job performance and employee wellness.
- The cohort used in this work includes over 100 individuals working as full-time nursing professions and 84 individuals worked the day shift.
- This work focuses on analyzing PPG data collected using Fitbit Charge 2.

Table 1: Average/Std of valid PPG recording length in hours in TILES study

<table>
<thead>
<tr>
<th></th>
<th>All Nurses</th>
<th>Day-shift</th>
<th>Night-shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1162.63</td>
<td>1182.50</td>
<td>1131.30</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>403.48</td>
<td>401.51</td>
<td>404.60</td>
</tr>
</tbody>
</table>

Table 2: Prediction accuracy of work status using motif derived features

<table>
<thead>
<tr>
<th></th>
<th>HIME Motif, $D$</th>
<th>Motif, $D$</th>
<th>Motif, $D$</th>
<th>Motif, $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>at pct=1%</td>
<td>65.20%</td>
<td>67.20%</td>
<td>68.35%</td>
<td></td>
</tr>
<tr>
<td>at pct=2%</td>
<td>66.50%</td>
<td>67.20%</td>
<td>68.54%</td>
<td></td>
</tr>
<tr>
<td>at pct=3%</td>
<td>66.62%</td>
<td>67.50%</td>
<td>68.54%</td>
<td></td>
</tr>
</tbody>
</table>

Thanks for Intelligence Advanced Research Projects Activity (IARPA) and MOSAIC program for their support.

Acknowledgement

May 10, 2019