Machine Learning to Assess Pilots’ Cognitive State
March 21, 2018

Tina Heinich
AST, Data Systems
Computer Engineer, OCIO Data Science Team
Summary

• Goal
• The Data
• Challenges
• Current Pipeline
• Future Work
The Team

• Tina Heinich (NASA)
• Angela Harrivel (NASA)
• Chad Stephens (NASA)
• Robert Milletich (Booz Allen Hamilton)
• Kellie Kennedy (NASA)
• James Comstock (NASA)
• Alan Pope (NASA Distinguished Research Associate)
• Charles Liles (NASA)
• Mary Carolyn Last (Analytical Mechanics Assoc.)
• Nijo Abraham (NASA)
• Nick Napoli (UVA)
• And many more....
In 2001-2010, Commercial Aviation Safety Team (CAST) identified unsafe cognitive states as a key factor in almost all studied airplane accidents.

Want to predict if a subject is currently in an unsafe cognitive state

Create a machine learning model that can reliably predict the cognitive state a subject is in using various physiological sensors

Specific Goal: Using previously collected data in a non-flight scenario, can we predict cognitive states in an actual flight simulation?
Cognitive States

- Startle/Surprise (SS)
  - Includes the “onset” as well as the initial 13 seconds “recovery” afterwards

- Channelized Attention (CA)
  - Subject is focused on one particular task/input to the exclusion of all other tasks/inputs

- Low Workload (LW)
  - Demand requires minimal resources to complete

- “Other”
  - Not really a class, but whenever a pilot is NOT experiencing these states
Physiological Modalities

- 20-channel Electroencephalography (EEG)
  - ABM B Alert X24
- Galvanic Skin Response (GSR)
  - NeXus-10
- Electrocardiogram (EKG)
  - NeXus-10
- Respiration (R)
  - NeXus-10
Data was collected for 13 crews of pilot and copilot (26 total participants)
  • Because of various data collection issues, we concentrate primarily on 5 full crews (10 pilots)

Benchmark Tasks
  • 6 minute tasks that induce various cognitive states
  • Tasks were done on two different days
    • Known as “Day 1 Benchmark” and “Day 2 Benchmark”

Line-Oriented Flight Training (LOFT) Data
  • Flight scenario with full flight
Tour of the Data

The Features

- **EEG**
  - Focus on MVP Channels
  - Summary Statistics:
    - Quantiles: Q005, Q995, Q75 and Inter-Quantile Range
    - Skew and Kurtosis
    - Variance
  - 4 Power Band Features
    - Alpha Band Mean (4-8 Hz)
    - Beta Band Mean (8-13 Hz)
    - Theta Band Mean (13-22 Hz)
  - Engagement Index and Task Load Index

- **ECG**
  - Heart Rate Variability Mean and Variance
  - Summary Statistics
    - Quantiles: q005 and IQR
    - Skew and Kurtosis
    - Variance

- **GSR**
  - Avg Slope and Drop Score
    - “drop score” essentially counts how often the slope drops under a given threshold
  - Summary Statistics
    - Quantiles: q975
    - Variance
    - Skew

- **Respiration**
  - Respiration Rate
  - Summary Statistics
    - Quantiles q75, q975, and q025
    - Skew and Kurtosis
    - Variance

Data was divided into sliding 5 second windows with 1 second stride. Within each window, features for each modality were calculated.
Tour of the Data

Various Notes

- There is a class imbalance between SS, CA, and LW
- Class labels are largely based on state induction (not actual cognitive state)
- We can’t rely on the sensor data always being there
  - Sometimes a sensor falls off, or gets skewed a little
- There is often variance between the different datasets

<table>
<thead>
<tr>
<th></th>
<th>Bench1</th>
<th>LOFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA Windows</td>
<td>2852</td>
<td>2932</td>
</tr>
<tr>
<td>LW Windows</td>
<td>2841</td>
<td>6669</td>
</tr>
<tr>
<td>SS Windows</td>
<td>241</td>
<td>282</td>
</tr>
<tr>
<td>Other Windows</td>
<td>3504</td>
<td>31173</td>
</tr>
</tbody>
</table>

Future Work Note: Currently have experts and flight instructor reviews of the data recordings so to improve the class labels/find other cognitive states in the data.
• Found that a model built on Day 1 Bench won’t necessarily work on Day 2 Bench
Variance Challenge
Differences between Pilots

• Hard to predict from pilot to pilot; variance is more pronounced in certain modalities
Variance Challenge
Variance Challenge

Preferred Modalities

SS = Startle/Surprise
CA = Channelized Attention
LW = Low Workload
Variance Challenge

Preferred Modalities
Variance Challenge
Difference Between LOFT and Benchmarks

- Often difficult to translate between benchmark and LOFT
Variance Challenge
• Input raw sensor data
  • SS, LW, and CA Day 1 Benchmarks
  • 66% holdout of LOFT
• Generate features
• Create X State Detector for each state
• Each detector is comprised of multiple anomaly/novelty detectors (with the state being the anomaly) each built with a different modality combination
• Each anomaly detector is trained with benchmark data and individually tuned on the LOFT holdout
The Pipeline

Testing

- Deploy on remaining 34% of LOFT
- For each testing window:
  - Generate features
  - For each state detector:
    - Run data through each model in the ensemble and gather predictions
    - Hold a simple majority vote to choose the prediction
    - Alternative: return the probability:
      \[
      \frac{\text{number of yes votes}}{\text{number of models}}
      \]
### Results

#### SS Detector Metrics

<table>
<thead>
<tr>
<th>Pilot</th>
<th>AUC</th>
<th>ACC</th>
<th>SS ACC</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>0.5073</td>
<td>0.9029</td>
<td>0.0909</td>
<td>0.9236</td>
</tr>
<tr>
<td>14</td>
<td>0.6309</td>
<td>0.8849</td>
<td>0.3636</td>
<td>0.8981</td>
</tr>
<tr>
<td>17</td>
<td>0.7504</td>
<td>0.5099</td>
<td>1</td>
<td>0.5008</td>
</tr>
<tr>
<td>18</td>
<td>0.5495</td>
<td>0.8162</td>
<td>0.2727</td>
<td>0.8263</td>
</tr>
<tr>
<td>21</td>
<td>0.8945</td>
<td>0.7929</td>
<td>1</td>
<td>0.7889</td>
</tr>
<tr>
<td>22</td>
<td>0.6772</td>
<td>0.7165</td>
<td>0.6364</td>
<td>0.718</td>
</tr>
<tr>
<td>23</td>
<td>0.8752</td>
<td>0.9283</td>
<td>0.8182</td>
<td>0.9323</td>
</tr>
<tr>
<td>24</td>
<td>0.662</td>
<td>0.6012</td>
<td>0.7273</td>
<td>0.5968</td>
</tr>
<tr>
<td>25</td>
<td>0.7034</td>
<td>0.6812</td>
<td>0.7273</td>
<td>0.6794</td>
</tr>
<tr>
<td>26</td>
<td>0.7384</td>
<td>0.6644</td>
<td>0.8182</td>
<td>0.6585</td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td><strong>0.6989</strong></td>
<td><strong>0.7499</strong></td>
<td><strong>0.6455</strong></td>
<td><strong>0.7523</strong></td>
</tr>
<tr>
<td><strong>STD</strong></td>
<td><strong>0.1179</strong></td>
<td><strong>0.1314</strong></td>
<td><strong>0.2916</strong></td>
<td><strong>0.1388</strong></td>
</tr>
</tbody>
</table>

#### CA Detector Metrics

<table>
<thead>
<tr>
<th>Pilot</th>
<th>AUC</th>
<th>ACC</th>
<th>CA ACC</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>0.4991</td>
<td>0.5643</td>
<td>0.3554</td>
<td>0.6429</td>
</tr>
<tr>
<td>14</td>
<td>0.6321</td>
<td>0.754</td>
<td>0.3636</td>
<td>0.9006</td>
</tr>
<tr>
<td>17</td>
<td>0.4554</td>
<td>0.4189</td>
<td>0.51</td>
<td>0.4008</td>
</tr>
<tr>
<td>18</td>
<td>0.5291</td>
<td>0.3278</td>
<td>0.83</td>
<td>0.2282</td>
</tr>
<tr>
<td>21</td>
<td>0.6438</td>
<td>0.7606</td>
<td>0.404</td>
<td>0.8836</td>
</tr>
<tr>
<td>22</td>
<td>0.6388</td>
<td>0.5271</td>
<td>0.8676</td>
<td>0.41</td>
</tr>
<tr>
<td>23</td>
<td>0.4617</td>
<td>0.3531</td>
<td>0.8091</td>
<td>0.1143</td>
</tr>
<tr>
<td>24</td>
<td>0.5771</td>
<td>0.675</td>
<td>0.2636</td>
<td>0.8905</td>
</tr>
<tr>
<td>25</td>
<td>0.6354</td>
<td>0.5336</td>
<td>0.8378</td>
<td>0.433</td>
</tr>
<tr>
<td>26</td>
<td>0.4448</td>
<td>0.349</td>
<td>0.6351</td>
<td>0.2545</td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td><strong>0.5517</strong></td>
<td><strong>0.5263</strong></td>
<td><strong>0.5876</strong></td>
<td><strong>0.5158</strong></td>
</tr>
<tr>
<td><strong>STD</strong></td>
<td><strong>0.0789</strong></td>
<td><strong>0.1558</strong></td>
<td><strong>0.2236</strong></td>
<td><strong>0.2799</strong></td>
</tr>
</tbody>
</table>

**AUC:** Receiver Operating Characteristic Area Under the Curve  
**ACC:** Straight Accuracy Metric  
**X ACC:** How well did we predict the state?  
**Other:** How well did we predict “Other” as “Other”?
Takeaways and Future Work

• Generalization is difficult (in general 😊)
  • Some modalities are easier to generalize than others
  • Some state detection is easier to generalize than others
  • Some machine learning tools can be used to mitigate some of this

• Work is continuing on many avenues:
  • Different Features
  • Different Modalities
  • Different Sensors
  • Concentrating on Startle/Surprise
Questions?


• Li, F., “Improving Engagement Assessment by Model Individualization and Deep Learning,” Dissertation, Old Dominion University, 2015.


• Bao, Forrest Sheng. “PyEEG.” PyEEG Reference Guide, 0.02 r1, pyeeg.sourceforge.net/.


• A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, M. Hämäläinen, MNE software for processing MEG and EEG data, NeuroImage, Volume 86, 1 February 2014, Pages 446-460, ISSN 1053-8119