Assessing Multiple Imputation of Missing Values for Robust Analysis of Telehealth Kiosk Data

Hava Chaptoukaev, Maxime Beurey, Juliette Raffort, Maria A. Zuluaga

IA et santé : approches interdisciplinaires
June, 29th 2022
Introduction

• **Telehealth** allows the distribution of health-related services.

• Promising avenue for **prevention**, and **remote diagnosis** and **monitoring** of diseases.

• Can be a solution when **access** to care is **restricted**.
Bodyo AiPod

• Stand-alone telehealth kiosk.

• Measures 27 health indicators in 6 minutes.

• 4 sensors collect information:
  • a scale
  • a body composition sensor
  • an oximeter
  • a blood pressure sensor
Deploying the AiPod

• In non-clinical contexts **sensors may fail**, leading to incomplete data.

• If one sensor fails **all measures** collected by the sensor go missing.

• We cannot afford to **discard** incomplete observations.

**Problem:**
How to deal with missing data?
Working with missing data

• We investigate two ways to deal with missing data:
  • Imputation schemes to fill the missing values.
  • A set approach that avoids imputation.
Imputation of missing values
Imputation schemes

- We still want to keep observations with missing values.
- Imputation schemes allow to fill missing values.
- Imputation needs to preserve the integrity of the original data.
Multiple Imputation with Denoising Autoencoders (MIDA)

- The MIDA architecture[1] imputes missing values.
- Based on a denoising autoencoder.
- MIDA masker not suitable for our problem.

**Our Approach:**
Modify the masker to imitate the pattern of a failing sensor.

Evaluation: feature reconstruction error

Average Mida MAE = 15.893
Mean MAE = 40.249
Evaluation: feature distribution
Evaluation : blood pressure classification

• Dataset: **329 samples** with **24 features**.

• Data from at least one of the sensors is missing for **48 samples**.

• **Use case**: assess if imputation improves the binary classification of BP categories according to 2 categories.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>F1-score</th>
<th>Precision</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.67</td>
<td>0.65</td>
<td>0.62</td>
<td>0.69</td>
</tr>
<tr>
<td>Mean</td>
<td>0.64</td>
<td>0.67</td>
<td>0.59</td>
<td><strong>0.77</strong></td>
</tr>
<tr>
<td>Our method</td>
<td><strong>0.71</strong></td>
<td><strong>0.71</strong></td>
<td><strong>0.65</strong></td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>
Limitations of imputation

• There may be additional information, not collected by the sensors.

• In our dataset, only 105 samples (out of 329) have no missing values.

• Imputation of poorly represented information can introduce significant biases in the learning process.
Set approach to learn with missing values
Set models

• Most classical machine learning models require fixed-dimensional inputs.

• Sets allow to overcome this limitation.

• Good alternative to learn with missing values.
Set models

• **Idea:** use permutation invariant neural networks.

• **Permutation invariant function:** indifferent to the ordering of its input.

**Theorem 1.** A function $f$ operating on a set $X$ having elements in a countable universe, is a valid set function iff there are functions $\varphi : R \to Z$ and $\rho : Z \to R$ such that

$$f(X) = \rho \left( \sum_{x \in X} \varphi(x) \right)$$

(1)
Deep Sets architecture

Input

\[ X \in \mathbb{R}^M \]

\[ \phi \]

\[ \phi(x_1), \phi(x_M) \]

\[ \mathbb{R}^{N \times M} \]

\[ + \]

\[ \mathbb{R}^N \]

Output

\[ f(x_1, \ldots, x_M) \]

\[ \mathbb{R} \]

Our approach

• Each input vector $X_i$ is encoded as a set of permutation invariant observations $x_j$.
• Each $x_j$ is represented as a tuple $(v_j, m_j)$ such that:
  $$X_i := \{(v_1, m_1),\ldots,(v_p, m_p)\}$$
• The whole dataset can then be described as:
  $$D := \{(X_1, y_1),\ldots,(X_n, y_n)\}$$

Our Approach:
Allows us to deal with missing values.
Evaluation: diabetes classification

- The Pima Indians Diabetes\cite{3} database is composed of 768 samples and 8 features.
- Up to 374 samples have missing values across 5 features.

![Bar chart showing the percentage of missing values across different features.](https://www.kaggle.com/competitions/diabetes-classification/data)

\cite{3} https://www.kaggle.com/competitions/diabetes-classification/data
Evaluation: benchmark

- **Benchmark**: Logistic regression, Random Forests and Gradient Boosting.

- Missing values need to be *imputed* first for the benchmark.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F1-score</th>
<th>Precision</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean imp. + LR</td>
<td>0.753</td>
<td>0.61</td>
<td>0.70</td>
<td>0.52</td>
</tr>
<tr>
<td>Mean imp. + RF</td>
<td>0.772</td>
<td>0.65</td>
<td>0.71</td>
<td>0.59</td>
</tr>
<tr>
<td>Mean imp. + GB</td>
<td>0.727</td>
<td>0.59</td>
<td>0.62</td>
<td>0.56</td>
</tr>
<tr>
<td>Our method</td>
<td>0.792</td>
<td>0.71</td>
<td>0.68</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Concluding remarks

• The problem of missing values is a particularly sensitive issue in the medical field.

• We proposed two simple yet robust models that yield good performances.

• Imputation methods should be used sparingly to avoid biases in the learning.
Ongoing work

• Develop a way to compute a **weighted aggregation**.

• Test the method on the **AiPod data**.

• Investigate the **combination** of the two approaches.
Assessing Multiple Imputation of Missing Values for Robust Analysis of Telehealth Kiosk Data

Hava Chaptoukaev, Maxime Beurey, Juliette Raffort, Maria A. Zuluaga

IA et santé : approches interdisciplinaires