Evaluation of machine learning-readiness in Alzheimer’s disease data cohorts

Vijaya B. Kolachalama, PhD

Associate Professor,
Department of Medicine, Boston University School of Medicine
Department of Computer Science, Boston University
Founding Member, Faculty of Computing & Data Sciences, Boston University
Web: http://sites.bu.edu/vkola; Twitter: @vkola_lab
Introduction

• Build assistive tools for neurology clinics
  – **Team**: Shangran Qiu, Prajakta Joshi, Xiao Zhou, Matthew Miller, Chonghua Xue, Michael Romano, Cody Karjadi, Joyce Lee, Akshara Balachandra, Diala Lteif, Sandeep Sreerama, Caitlin Newman, Yi Zheng, Lindsey Claus, Yichi Zhang, Olivia Zhou, Rushin Gindra, Anika Walia, Aakash Bhatnagar, Meagan Lauber, Meysam Ahagaran

• PhD students and postdocs in computer science
• MD students, residents and clinical fellows at BUSM
• Trainees with background in neuroscience
• Background in more than a single discipline
Introduction

Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification

Shangran Qiu, Prajakta S Joshi, Matthew I Miller, Chonghua Xue, Xiaozhu Zhou, Cody Karjadi, Gary H Chang, Anant S Joshi, Brigid Dwyer, Shuhan Zhu, Michelle Kaku, Yan Zhou, Yazan J Alderazi, Arun Swaminathan, Sachin Kedar, Marie-Helene Saint-Hilaire, Sanford H Auerbach, Jing Yuan, E Alton Sartor, Rhoda Au, Vijaya B Kolachalam

Nature Communications

Multimodal deep learning for Alzheimer’s disease dementia assessment

Shangran Qiu, Matthew I Miller, Prajakta S Joshi, Joyce C Lee, Chonghua Xue, Yunruo Ni, Yuwei Wang, Ileana De Anda-Duran, Phillip H Hwang, Justin A Cramer, Brigid C Dwyer, Honglin Hao, Michelle C Kaku, Sachin Kedar, Peter H Lee, Asim Z Mian, Daniel L Murman, Sarah O’Shea, Aaron B Paul, Marie-Helene Saint-Hilaire, E Alton Sartor, Aneeta R Saxena, Ludy C Shih, Juan E Small, ... Vijaya B Kolachalam

Nature Communications 13, Article number: 3404 (2022)
Introduction

- **Machine learning:** data + model → prediction
- Quality of prediction depends on both good **models** and high-quality **data**
- Scientists spend about \(\frac{3}{4}\) of their time in iterative pre-processing of data
- Pre-processing: **cleansing**, **validation** and **transformation**
Data quality measures

- **Class Overlap**: Overlapping regions among different classes
- **Label Purity**: Noise ratio and the number of noisy samples in the data
- **Class Parity**: Class imbalance ratio
- **Feature Relevance**: Importance of each feature with respect to the target variable (class) and other features
- **Data Homogeneity**: Transformation of data into the user's intended format
- **Data Fairness**: Identifies the bias in the dataset and returns the disparate impact score
- **Correlation Detection**: Detect correlated features
- **Data Completeness**: Detects missing values
- **Outlier Detection**: Detect and remove outliers (deviates so much from other observations) in the dataset
- **Data Duplicates**: Remove duplicate records to clean the data
Assumptions

- Class Overlap:
- Label Purity:
- Class Parity:
- Feature Relevance:
- Data Homogeneity:
- Data Fairness:
- Correlation Detection:
- DataCompleteness:
- Outlier Detection:
- Data Duplicates:

Most people evaluate these quality metrics independently and assume that poor quality affects outcomes.

Framework: Evaluate the data quality by defining an objective and looking at the overall model performance.
Data readiness in AD

- Alzheimer’s (ADNI)
- Explainable Data
- Measurements
  - Outliers
  - Missing values
  - Correlated features
  - Noisy data

1. Unstructured Data
2. Determine Level of Readiness
3. Remediation and Cleansing Data
4. Structured Data
5. Machine Learning
6. Test Data
7. Evaluation
8. Prediction

Graph
Table
Time Series
Proposed method

Data Space $D$

Dataset $D$ [NxM]

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>$c_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sub-table $T_1$ [N1 x M1]
Sub-table $T_2$ [N2 x M2]
$d$ Sub-tables
Sub-table $T_d$ [Nd x Md]

Dataset Stratification

Data Readiness Space $D'$

$k$ Data quality features
$F = \{f_1, f_2, ..., f_k\}$

Dataset $D'$ [dxk]

Readiness Evaluation

Supervised Learning
Unsupervised Learning
Learning weights of data quality measures

Sub-tables

$T_1 \quad N_1 \times M_1$

$T_2 \quad N_2 \times M_2$

$\cdots$

$T_d \quad N_d \times M_d$

Accuracy of classification

Classification

Clustering

Accuracy of Clustering

Sub-tables $f_1 \quad f_2 \quad \cdots \quad f_k$

<table>
<thead>
<tr>
<th>Sub-tables</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$\cdots$</th>
<th>$a_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_d$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weight Vector $W = (w_1, \ldots, w_k)$

Calculating weighted total quality of sub-tables

Regression
Problem’s computational complexity

NP-Complete problem similar with *Subset-Sum* Problem

<table>
<thead>
<tr>
<th>Dataset $D$</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>$c_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data $N$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Running the proposed algorithm

Original Dataset $D$ [$N \times M$]

<table>
<thead>
<tr>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$...$</th>
<th>$c_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$...$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data $N$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Running of the algorithm

Datasets $D'$ [$d \times k$]

- **Run 1**
  - $D'_1$
  - $f_1$
  - $f_2$
  - $...$
  - $f_k$
  - $T_1 = [N_1 \times M_1]$
  - $T_2 = [N_2 \times M_2]$
  - $...$
  - $T_d = [N_d \times M_d]$

- **Run 2**
  - $D'_2$
  - $f_1$
  - $f_2$
  - $...$
  - $f_k$
  - $T_1 = [N_1 \times M_1]$
  - $T_2 = [N_2 \times M_2]$
  - $...$
  - $T_d = [N_d \times M_d]$

- **Run $R$**
  - $D'_R$
  - $f_1$
  - $f_2$
  - $...$
  - $f_k$
  - $T_1 = [N_1 \times M_1]$
  - $T_2 = [N_2 \times M_2]$
  - $...$
  - $T_d = [N_d \times M_d]$

Local Optimums

- $T_1^*, \tilde{f}_1^*$
- $T_2^*, \tilde{f}_2^*$
- $T_R^*, \tilde{f}_R^*$

Global Optimum $T^*, \tilde{f}^*$
### Static features of the ADNI dataset (ADNIMERGE table)

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Column Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines (44 features)</strong></td>
<td>CDRSB_bl, ADAS11_bl, ADAS13_bl, ADASQ4_bl, MMSE_bl, RAVLT_immediate_bl, RAVLT_learning_bl, RAVLT_forgetting_bl, RAVLT_perc_forgetting_bl, LDELTOTAL_BL, DIGITSCOR_bl, TRABSCOR_bl, FAQ_bl, mPACCdigit_bl, mPACCtrailsB_bl, Ventricles_bl, Hippocampus_bl, WholeBrain_bl, Entorhinal_bl, Fusiform_bl, MidTemp_bl, ICV_bl, MOCA_bl, EcogPtMem_bl, EcogPtLang_bl, EcogPtVisspat_bl, EcogPtPlan_bl, EcogPtOrgan_bl, EcogPtDivatt_bl, EcogPtTotal_bl, EcogSPMem_bl, EcogSPLang_bl, EcogSPVisspat_bl, EcogSPPlan_bl, EcogSPOrgan_bl, EcogSPDivatt_bl, EcogSPTotal_bl, ABETA_bl, TAU_bl, PTAU_bl, FDG_bl, PIB_bl, AV45_bl, FBB_bl.</td>
</tr>
<tr>
<td><strong>Demographics (6 features)</strong></td>
<td>AGE, PTGENDER, PTEDUCAT, PTETHCAT, PTRACCAT, PTMARRY</td>
</tr>
<tr>
<td><strong>Genetics (one feature)</strong></td>
<td>APOE4</td>
</tr>
<tr>
<td><strong>Diagnosis class (one feature)</strong></td>
<td>DX_bl</td>
</tr>
</tbody>
</table>
Simulation

- Random sub-tables were generated ($R=4$, $d=500$)
- Five diagnosis groups of patients at baseline (CN, pMCI, sMCI, and AD)
- Data quality features: Pearson Correlation (PC), Missing Values, Spearman Correlation, Outliers, and Class overlap
- Classification method: Random Forest (RF) classifier
- Clustering method: Agglomerative and k-means clustering (Silhouette Coefficient for calculating the accuracy of clustering)
- Random Forest Regression for learning the weight vector
Results

- **Mean weight vector:** $W^* = (0.2005, 0.207, 0.2812, 0.0115, 0.2998)$ for PC, Spearman Correlation, Missing Values, Outliers, and Class Overlap, respectively

- Weights of five data quality measures:

<table>
<thead>
<tr>
<th>Run</th>
<th>PC</th>
<th>Spearman</th>
<th>Missing Values</th>
<th>Outliers</th>
<th>Class Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1727</td>
<td>0.2331</td>
<td>0.2729</td>
<td>0.0088</td>
<td>0.3125</td>
</tr>
<tr>
<td>2</td>
<td>0.2123</td>
<td>0.181</td>
<td>0.3034</td>
<td>0.0042</td>
<td>0.299</td>
</tr>
<tr>
<td>3</td>
<td>0.2062</td>
<td>0.2177</td>
<td>0.2733</td>
<td>0.0132</td>
<td>0.2895</td>
</tr>
<tr>
<td>4</td>
<td>0.2109</td>
<td>0.1962</td>
<td>0.275</td>
<td>0.0197</td>
<td>0.2982</td>
</tr>
</tbody>
</table>

- Details of the best sub-table:

<table>
<thead>
<tr>
<th>Run</th>
<th>Rows</th>
<th>Columns</th>
<th>PC</th>
<th>Spearman</th>
<th>Missing values</th>
<th>Outliers</th>
<th>Class Overlap</th>
<th>Classification Accuracy</th>
<th>Total Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1532</td>
<td>21</td>
<td>0.6708</td>
<td>0.6736</td>
<td>0.7685</td>
<td>1</td>
<td>0.495</td>
<td>0.7383</td>
<td>0.6499</td>
</tr>
<tr>
<td>2</td>
<td>1394</td>
<td>18</td>
<td>0.6563</td>
<td>0.6502</td>
<td>0.7277</td>
<td>1</td>
<td>0.4924</td>
<td>0.7188</td>
<td>0.6299</td>
</tr>
<tr>
<td>3</td>
<td>1474</td>
<td>16</td>
<td>0.7606</td>
<td>0.7492</td>
<td>0.8023</td>
<td>1</td>
<td>0.3212</td>
<td>0.6459</td>
<td>0.641</td>
</tr>
<tr>
<td>4</td>
<td>1994</td>
<td>16</td>
<td>0.7198</td>
<td>0.7163</td>
<td>0.6372</td>
<td>1</td>
<td>0.6433</td>
<td>0.7678</td>
<td>0.6761</td>
</tr>
</tbody>
</table>
Take home messages

• Data readiness is important to evaluate, and it encompasses various aspects related to data quality
• We developed a data readiness framework to evaluate AD cohorts with an objective to improve dementia assessment
• **Limitation:** Building frameworks to evaluate data readiness without considering objectives can reduce their value
• **Limitation:** Over reliance on such measures can lead to good modeling frameworks but can limit progress towards building true AI-based systems
Acknowledgments

Funding: