Machine Learning-Spring 2016
School of Computer and Information Sciences
Florida International University

Mining Big Neuron Morphological Data

MightyMorph Group:
Maryam Aghili
Maria Presa Reyes
Jingan Qu

April 27, 2016

Content

What is Neuromorpholog?

Studying

- Form of the neurons
- Shape of the neurons
- Geometry of the neurons

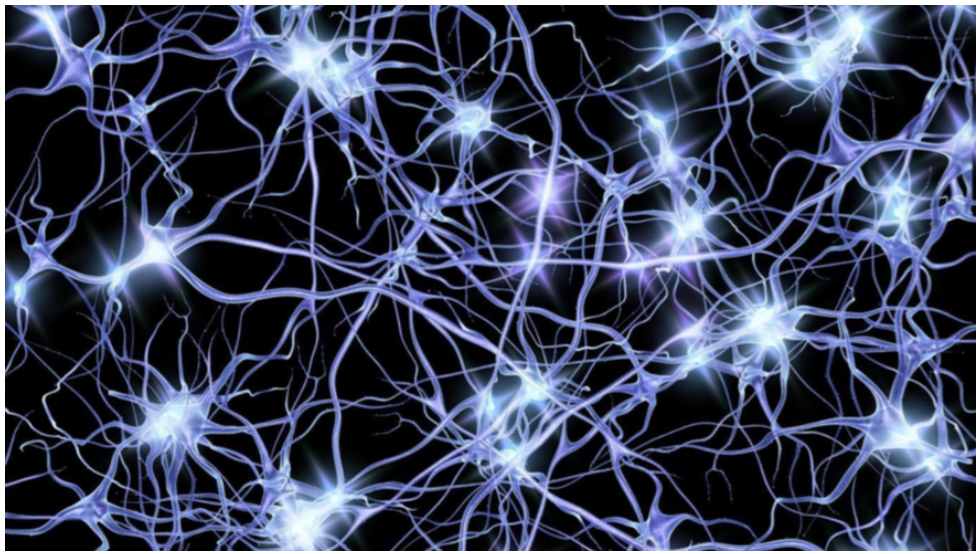
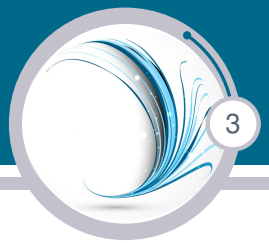
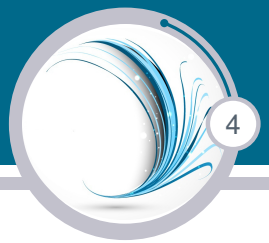


Figure: Neurons



Why it is important

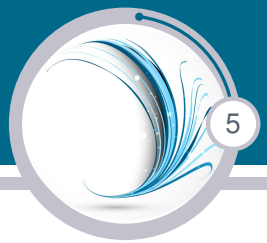
- Shape.
- Functionality



Also important for...

Tools for diagnosis the root of the neural diseases such as

- Emotional disorders
- Learning disabilities
- Malfunction of some parts in the injured brain



Goal

Propose a high performance algorithm with the highest accuracy that can classify tens of thousands of neurons into different cell types and brain regions using their morphological information.

Input

Digitally reconstructed neurons provided by the web page:
`NeuroMorpho.org`

Output

A comparison on the performance of each algorithm over this dataset.

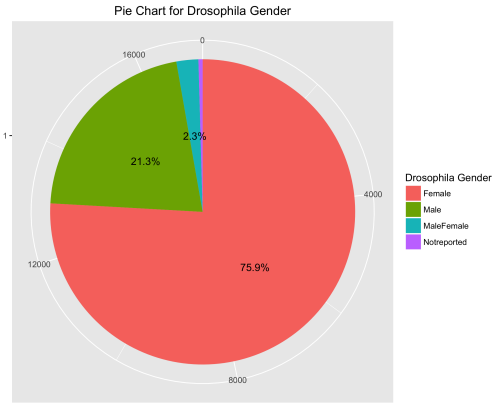
Introduction
General Approach

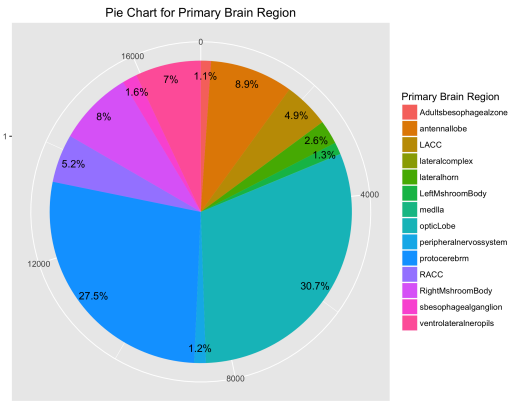
- Create feature ,label set of information with the help of L-Measure from the reference site
- Apply all the state of art machine learning algorithms on the data to have a baseline for our performance.
- Start applying different types of techniques to improve the baseline performance until reaching a higher level of performance.

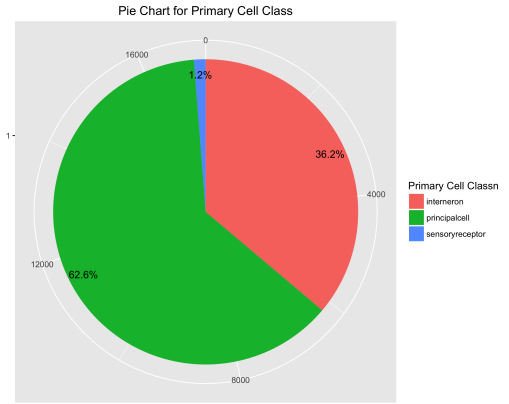
Preprocessing

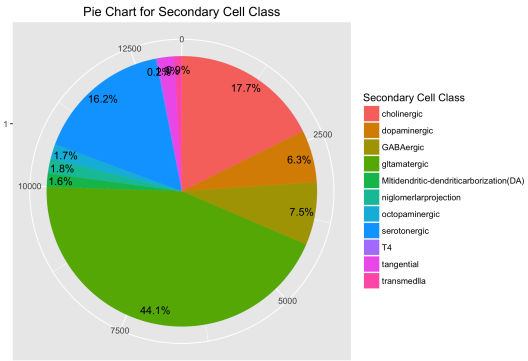
- Features Selection
- Dependent feature removal
- Redundant feature removal using correlation
- Useless feature removal without too much change
- Principle Component Analysis

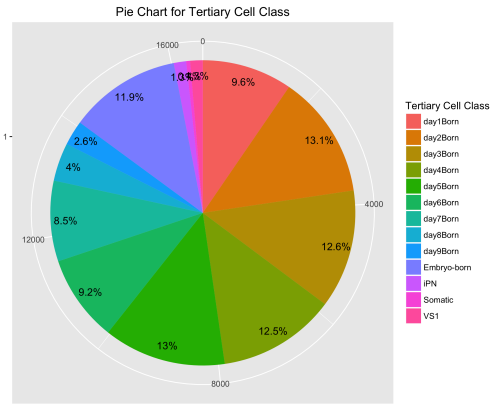
Introduction
Data Visualization











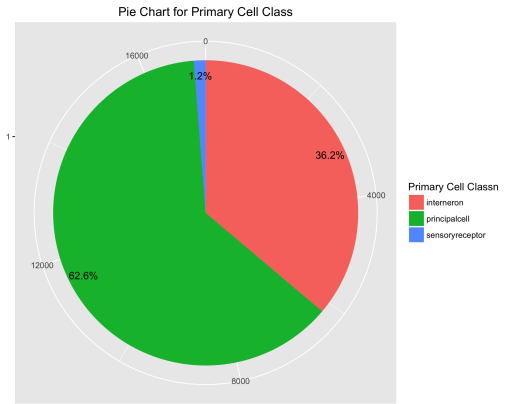


Figure: Pie Charts for all Label Variables

Method
Baseline Accuracy

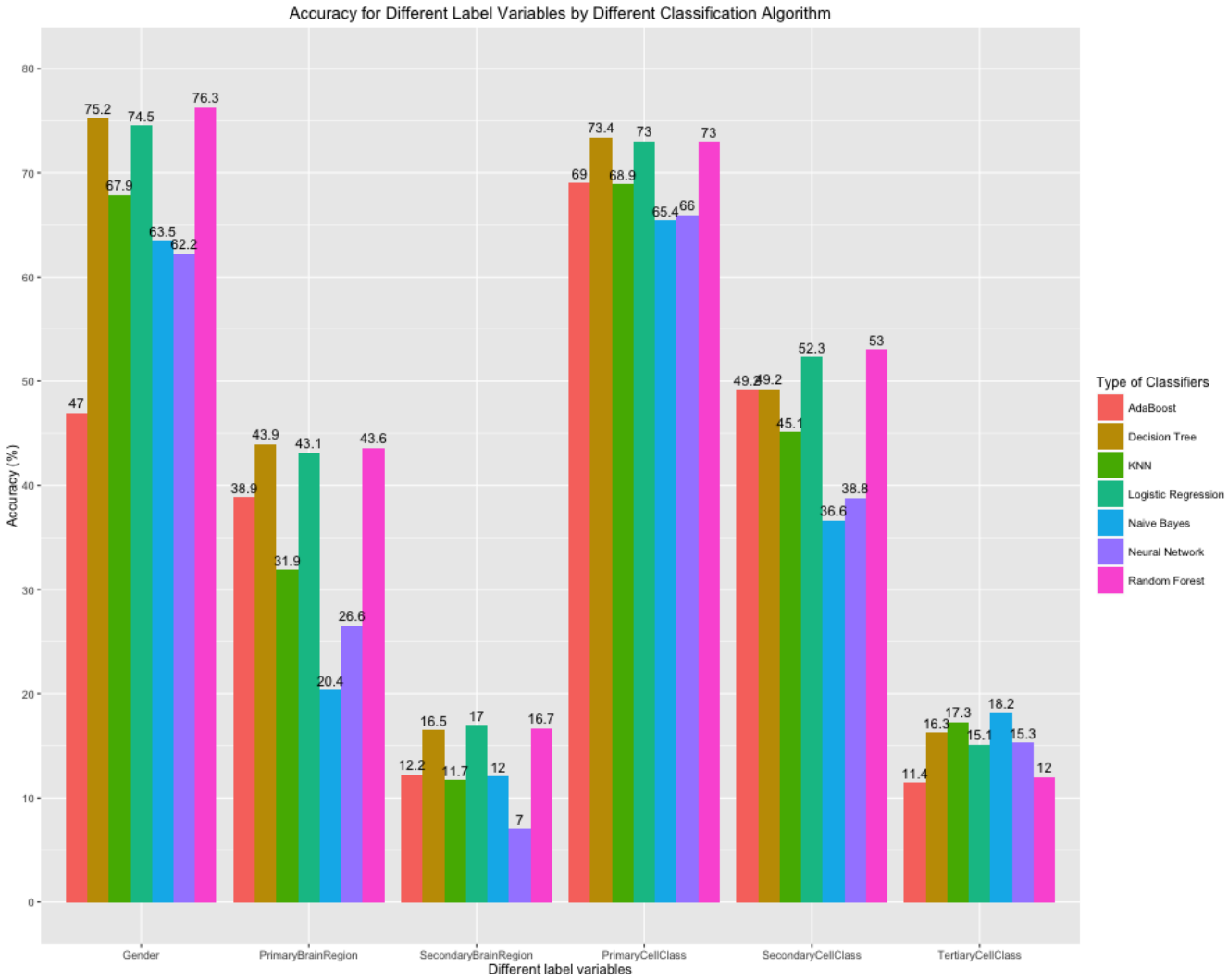
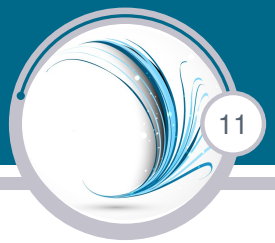


Figure: Baseline Accuracy for Different Label Variables by Different AL

Method
Data Preparation

- Mark the classes like "Unknown", "Notreported", etc. in label variables as missing value.
- Deal with missing value (Too many, about one fourth, put them in training data and make them up with prediction)
- Remove the classes in label variables with low rate of **occurrence**
- Remove the features that are constant.

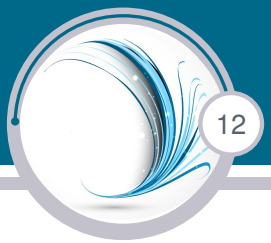


Missing Value

Make up the missing value by **prediction**

Label Variables Relationship

- Chi-square test to prove each two label variables have a significant relationship
- Make use of the information from other label variables.



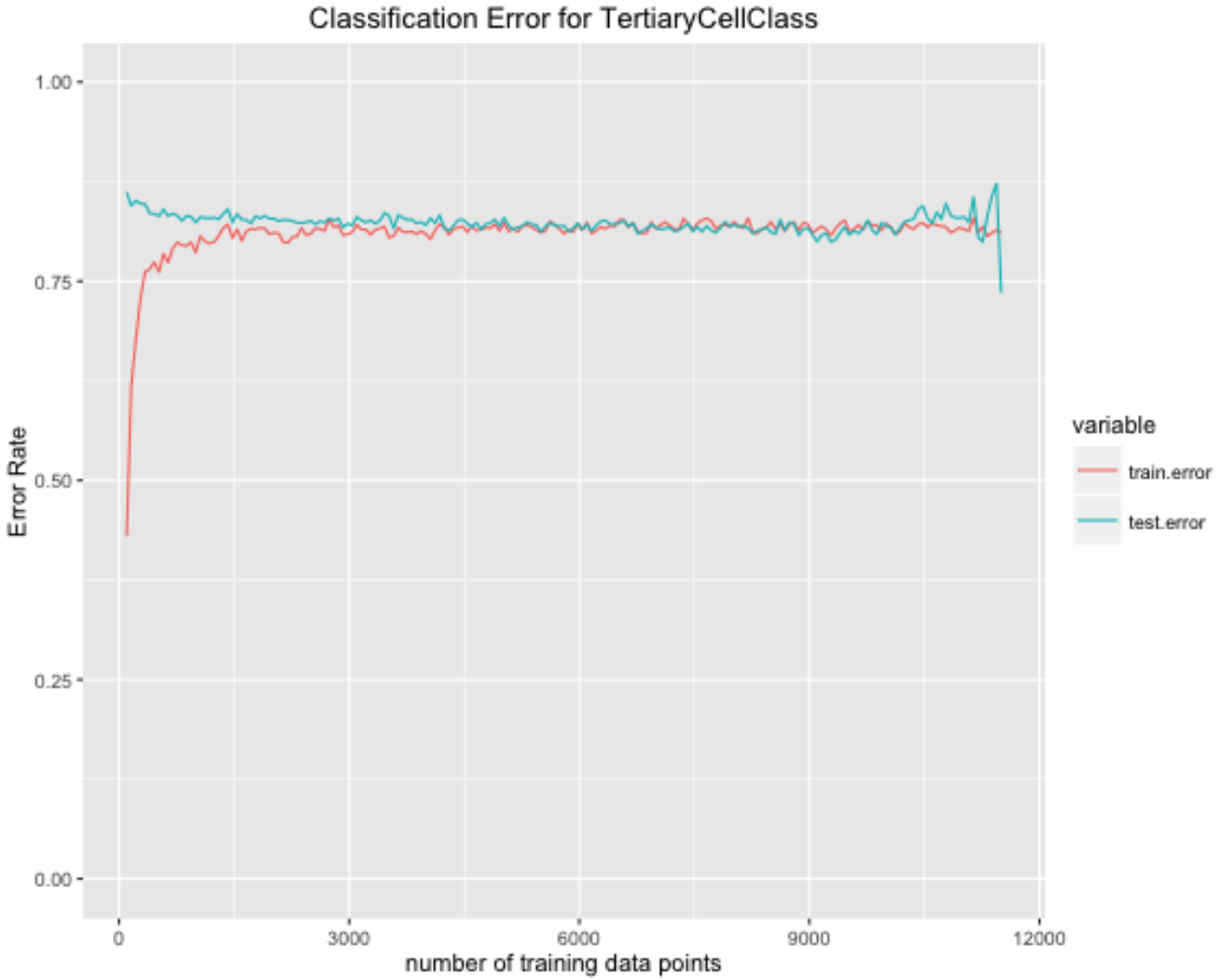


Figure: Learning Curve of Label Variable TertiaryCellClass

Ensemble Method

- Divide the training data into 3 folds, each fold has around 3000 records.
- For each fold, apply **KNN**, **Logistic Regression**, **Decision Tree**, **AdaBoost**, **Naive Bayes**, **Neural Network**, **Random Forest** separately to build 7 models to predict one label variable.
- Combine these 7x3 models into one final model by weighted ensemble method.
- The weights in ensemble are the baseline accuracies of each algorithms for each label variable

Experimental Results

Accuracy of Before & After

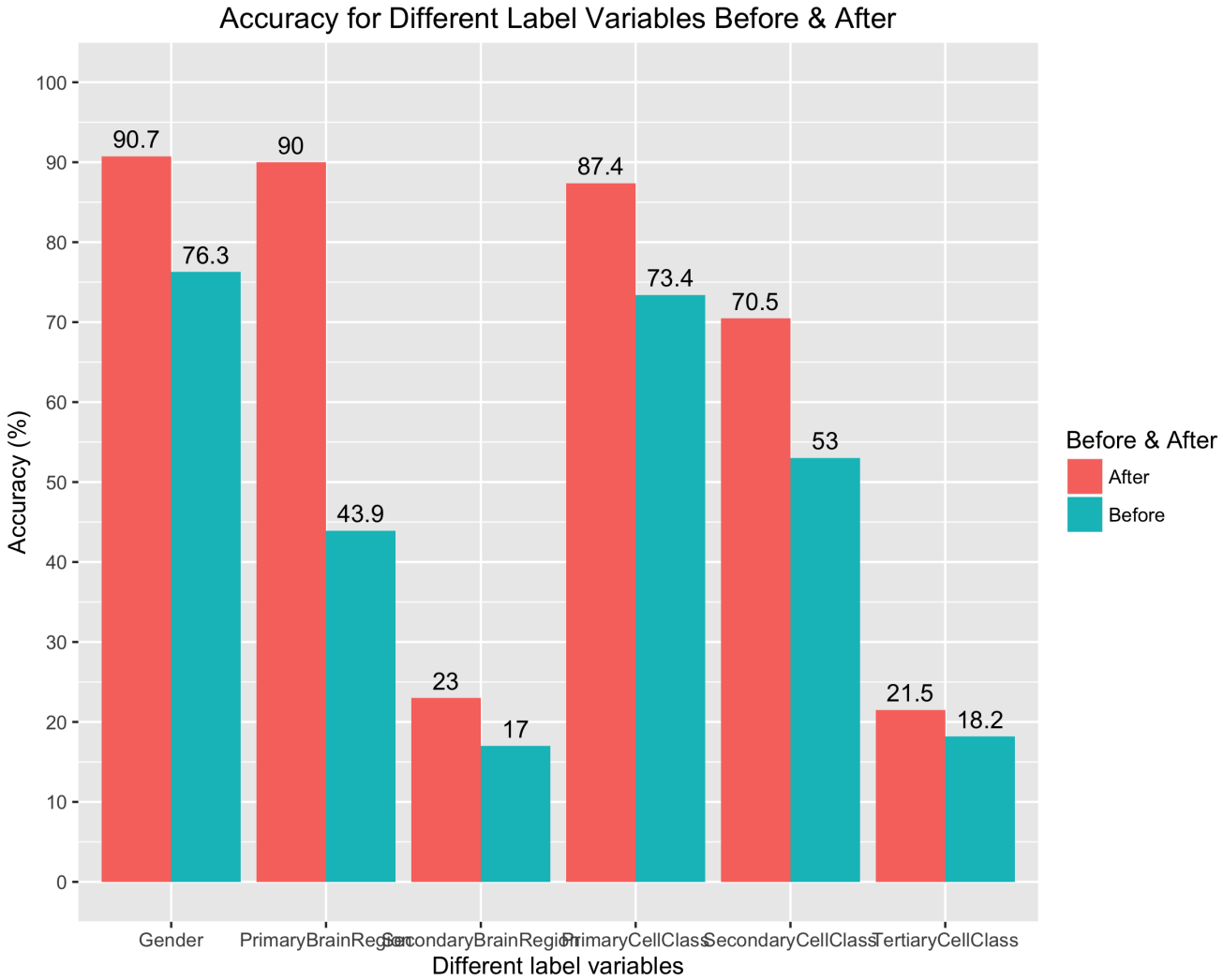


Figure: Accuracy for Different Label Variables Before & After

Discussion of Results

- The label variable with less classes, the accuracy improve more.
- The label variables like SecondaryBrainRegion and TertiaryCellClass, maybe not suitable for classification

Future Work

- Include the feature selection or feature engine in our model, find out which variable is most important factor for classification
- Apply deep learning model to classify the 3D neuron structure directly.


Thank You !
#53740257