CS6024 Mini-Project Proposal
Gene expression analysis for age prediction

Balakrishnan A (CS20B012), Sooraj Srinivasan (CS20B075)

Jan-May 2023

Introduction

The goal of our project is to use data analysis and machine learning tools to analyze genome-wide RNA-seq profiles of human dermal fibroblasts to predict the chronological age of the person. Through this procedure, we wish to determine **biomarkers for human aging**, by identifying genes that are especially important in accurate age prediction. Doing so may even help determine genes especially susceptible to aging and age-related diseases, which can be further studied and targeted by research on (preventing/identifying) late-onset diseases.

Problem statement

We intend to use the *Age prediction using machine learning*[1] dataset, as used by Fleischer et al., who tackled the task of building a better model for predicting chronological age. Using this dataset, we intend to achieve the following:

1. Identify a better-performing class of models and/or a new ensembling method that performs better on the task of age prediction
2. Use the model to identify key genes involved in prediction, if significant

We expect to isolate a few key genes that prove very helpful in prediction and compare those to existing genes believed to be linked closely to aging. We also aim to train a well-performing model and produce a saved trained copy, which can then be directly used for this task with the right data.

Workflow outline

We plan to approach the problem in three phases, described as follows:

1. Exploratory data analysis [1 week]
    - Analysis of the RNA-seq data, to look for correlations, insignificant features
    - Also closely linked to the next phase, as feature elimination may also be performed while fitting, determining the important features on the fly
2. Model fitting [1-2 weeks]

- Comparison of performances of multiple learning algorithms and ensemble methods, in terms of prediction accuracy, speed, and memory usage
- Further hyperparameter tuning if necessary, for fine-tuning

3. Analysis of results [1 week]
    - Analysis of prediction results, whether within similar bounds as the ones established by Fleischer et al.
    - Identification of genes highly important in modeling, and comparison of findings (if significant) to existing genes identified/believed to be linked to aging

Deliverables include Python scripts/notebooks, either for the entire process or for the different phases mentioned.

Paper critique

Fleischer et al. attempts to build a better model for chronological age prediction than existing ones that use various models such as deep neural networks and ElasticNet regressors. They approach this problem with the help of a unique classification ensemble. The ensemble consists of multiple LDA classifiers, each trained on a *bin* of ages. The bin width (number of classes per bin) is N, say, 20, and each bin is shifted from the previous by 1. During prediction, a vote is performed using the predictions of all the classifiers, and the majority is considered. Ties are broken by choosing the lowest age. This ensembling proves very robust, attaining a significantly better performance than most existing classifiers for this task.

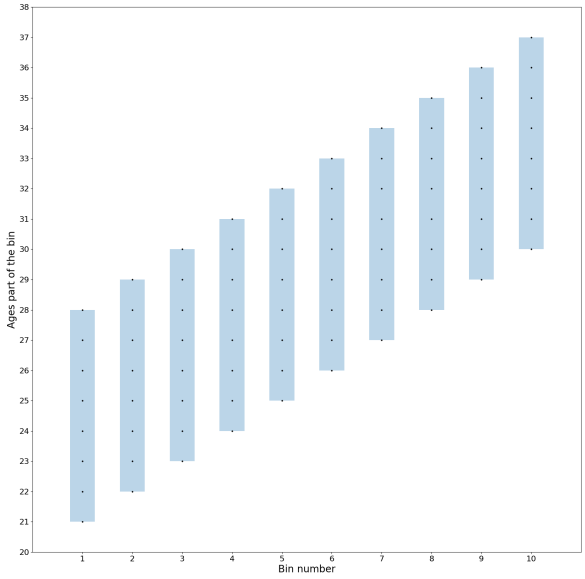

Figure 1: Bins of width 8 with a shift of 1. 10 bins are visualized, with the first bin starting at 21. The box describes the range of the bin, and the points detail the classes that the corresponding bin contains

Pros and cons

The robustness of the ensemble is also supported by its ability to detect rapid aging in individuals with HGPS, a syndrome believed to cause accelerated aging, leading to a dramatic shortening of the expected lifespan. The trained ensemble predicts ages significantly larger than the true ages, suggesting an ability to detect biological age and identify genetic biomarkers for aging and linked diseases.

A disadvantage of ensembling is difficulty in the interpretation of its results. This is true in the case of this ensemble method as well, making the analysis of key genes in the aging process difficult. As a result, this analysis is not presented in the paper.

The paper has used machine learning models like linear regression models, ElasticNet models, support vector machines, etc. A potential line of inquiry would be to attempt the classification using deep neural networks and see if the recent advances in deep learning provide better training and validation accuracy as compared to the ensemble model.

With regards to the implementation, the authors had chosen to write their code in Python 2.7.9[2] which has been sunsetted as of January 1, 2020. A modern implementation with Python 3.* is necessary if their work is to be reproduced.

References

[1] anup. Age prediction using machine learning, January 2019. URL https://doi.org/10.5281/zenodo.2545213.

[2] Jason G Fleischer. Predicting age from the transcriptome of dermal fibroblasts. source code v0.1. [internet]. github; 2018., December 2018. URL https://www.github.com/jasongfleischer/Predicting-age-from-the-transcriptome-of-human-dermal-fibroblasts/releases.

[3] Jason G Fleischer, Roberta Schulte, Hsiao H Tsai, Swati Tyagi, Arkaitz Ibarra, Maxim N Shokhirev, Ling Huang, Martin W Hetzer, and Saket Navlakha. Predicting age from the transcriptome of human dermal fibroblasts. *Genome biology*, 19:1–8, 2018.