Machine Learning for Polygenic risk score

Vu-Lam DANG
Universite Grenoble Alpes
Grenoble, France
vu-lam.dang@etu.univ-grenoble-alpes.fr

Supervised by: Dr. Michael G. B. Blum.

I understand what plagiarism entails and I declare that this report is my own, original work.
Vu-Lam DANG, 26/04/2019:

Abstract

The advent of genomic prediction as a viable diagnostic tool for certain disease have become a reality in the last decade. One particular interesting metric is called Polygenic Risk Score, which sumaries the genetic. For this internship project, we interested in applying advancement in the field of machine learning to improve the predictive power of PRS.

1 Introduction

1.1 Polygenic Risk Score

Polygenic Risk Score (PRS) [Dudbridge, 2013] is a single perimeter metric constructed from the weighted sum of associated alleles within each subject. Given a pair of traits $Y = (Y1, Y2)'$ 0 expressed as a weighted sum of m genetic effects and a bias indicate environmental and unaccounted genetic effects:

$$Y = \beta' G + \epsilon = [\sum_{i=1}^{m} \beta_{i1} G_1 + \epsilon_1, \sum_{i=1}^{m} \beta_{i2} G_2 + \epsilon_2]$$

where β is a $m * 2$ weight matrix, G is m length vector of genetic markers and ϵ is a pair of random error that's independent from G.

The Polygenic score is defined as:

$$\hat{S} = \sum_{i=1}^{m} \beta_{i1} G_1$$

Association between a trait and its composite score highly implies there exist a genetic signal among the markers, and the evident of genetic effect when there is no obvious candidate can be obtained. Currently polygenic score have been used for association testing rather than predicting complex traits [Dudbridge, 2013].

Howerver, [Inouye *et al.*, 2018] conducted a PRS study on 1.7 million genetic variants using UK Biobank database on Coronary Atery Disease, and found strong association between PRS and hazard ration for CAD; demonstrated the power of genomic risk prediction to stratify individuals, and highlights the possibility for genomic screening early in life to support risk predition and preventive treatement.

1.2 Regression Method

1.3 Objective

2 Proposal

For this work we would like to pursuit 2 different direction. The first direction is called stacking, in which one model with different values for input parameters, or several different models is combined to create more better prediction.

The second direction is to create a sparse model that closely resemble the predictive power of genome-wide risk score.

3 Methodology

4 Conclusion

Polygenic Score have been widely studied as a good tool to identify and early diagnose high risked individual. In order to make it a viable tool for clinical usage, improvement have to be make to decrease false negative and reduce computational complexity of the model. By incorporate machine learning technique, we hope to eventually develop PRS into a powerful and accurate metric and tool to assess and identify clinical risk factor.

References

[Dudbridge, 2013] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*, 9(3):1–17, 03 2013.

[Inouye *et al.*, 2018] Michael Inouye, Gad Abraham, Christopher P. Nelson, Angela M. Wood, Michael J. Sweeting, Frank Dudbridge, Florence Y. Lai, Stephen Kaptoge, Marta Brozynska, Tingting Wang, Shu Ye, Thomas R. Webb, Martin K. Rutter, Ioanna Tzoulaki, Riyaz S. Patel, Ruth J.F. Loos, Bernard Keavney, Harry Hemingway, John Thompson, Hugh Watkins, Panos Deloukas, Emanuele Di Angelantonio, Adam S. Butterworth, John Danesh, and Nilesh J. Samani. Genomic risk prediction of coronary artery disease in 480,000

adults. *Journal of the American College of Cardiology*, 72(16):1883–1893, 2018.