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 summaries 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 of 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 evidence of genetic effect when there is no obvious candidate can be obtained. Currently, PRS has been used for association testing rather than predicting complex traits [Dudbridge, 2013].
However, [Inouye *et al.*, 2018] conducted a PRS study on 1.7 million genetic variants using UK Biobank database on Coronary Artery Disease, and found a 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 prediction and preventive treatment.

1.2 Calculating β

The centrepiece of calculating polygenic score is to determine the weight matrix β.
Originally, β was calculated using Linear Regression. To some degree, this method provides good predictive power [Dudbridge, 2013]. However, this method overlooks some important genetic effects, such as Linkage Disequilibrium. Furthermore, all SNPs are used in the final calculation while not all are useful, lead to some loss in performance.
Clumping and Thresholding (C+T) is a derivative from simple Linear Regression PRS. In this method, only SNPs with p-values that are under a threshold are selected, and related SNPs are clumped together as a single input variable. We implemented this model in R using a package called *bigsnpr* [Privé *et al.*, 2018] that provide various functions for genomics computation.
To specifically combat Linkage Disequilibrium, LDPred provide a Python package and standalone application to calculate PRS, with LD modelled into the computation [Vilhjlmsson *et al.*, 2015].
lassosum [Mak *et al.*, 2017] is an R package that provides PRS using a Penalized Regression known as LASSO. This method is the state of the art, and so far provide the best AUC in all methods tested ($AUC = 0.73$).

1.3 Objective

In this work, we looking for methods to 1. enhance the prediction power of PRS and 2. reduce the computing power required to compute β and PRS. In particular, stacking and pruning techniques will be investigated to accomplish these goals.

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 are combined to create better predictions.
The second direction is to create a sparse model that closely resembles the prediction power of genome-wide risk score.

This method has the benefit of reducing computing complexity while (theoretically) maintaining prediction power.

3 Methodology

The models are trained on a sample dataset of 8000 data points, each with more than 600000 SNPs. A test dataset is also provided. The same dataset is also used to verify the result of other methods (described in section 2).
These datasets are provided by Florian Prive.
The computation on this work is written in R and Python. For R we specifically use 2 package bigstasr and bigsnpr, which provide methods to handle large dataset by using disk swap [Privé *et al.*, 2018].

4 Conclusion

Polygenic Score has been widely studied as a good tool to identify and early diagnose high risked individuals. In order to make it a viable tool for clinical usage, improvement has to be made to decrease false negative and reduce the computational complexity of the model. By incorporating 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.

[Mak *et al.*, 2017] Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469–480, 2017.

[Privé *et al.*, 2018] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael G.B. Blum. Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. *Bioinformatics*, 2018.

[Vilhjlmsson *et al.*, 2015] Bjarni J. Vilhjlmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindstrm, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, and et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576592, Oct 2015.