Approaches to Incorporating Whole Genome Information in the context of Mining Genetic Data

Siddharth Avadhanam
avadhana@msu.edu

1. PROBLEM DESCRIPTION

In this work I propose to evaluate classical whole genome regression approaches in the statistical genetics literature such as lasso and ridge-regression against newer ensemble methods that are being applied to data from Genome Wide Association Studies. I will follow this review paper by Scyzymczak et al. [1] and implement some of the more succesful approaches in a wheat data set. I will evaluate these procedures in terms of cross validation accuracy, statistical considerations and computational performance.

The high-dimensionality and complexity of data coming in from Genome Wide Association Studies poses a whole range of statistical and computational challenges in finding genetic covariates of important phenotypes (disease, yield , height etc). These data commonly contain thousands of individuals with information on hundreds of thousands of markers. [2] An important issue which has been given considerable attention in the statistical genetics literature is how to choose a subset of these markers so that the analysis of these data with standard commodity hardware and statistical methods becomes more tractable.[3] Methods which analyze a single-marker at a time are popular, but are severly limited in that they do not take into account interactions between markers, and come with their own set of challenges related to multiple testing.[4] It is thus becoming increasingly important to explore methods that incorporate information form the whole genome and possibly from outside the domain of traditional statistics. In this work I will take a first step in evaluating the popular approaches towards whole genome regression, and contrast it with newer ensemble methods and possibly (if time permits) deep learning approaches. I will evaluate these methods on a publicly available dataset for wheat, looking at statistical considerations such as robustness, cross-validation accuracy of prediction, and computational feasibility.

2. RELATED WORK

There is considerable literature in statistical learning approaches towards the analysis of GWAS data. Earlier work has touched upon single marker regression methods and the whole host of approaches that can be applied to deal with the multiple-testing issues. Penalized regression methods such as ridge regression,[5] variable selection methods such as the lasso and dimensionality reducing approaches such as principal components [6] have all been applied and studied extensively in the analysis of genome-wide data. Bayesian methods and mixed models approaches within specific communities such as animal breeding have been studied and applied for a long time, and are increasingly being used with genome-wide data. [3] The discovery of gene-gene interactions is an active area of research where machine learning tools are being extensively employed [7]

3. PROJECT MILESTONES

This section will detail milestones so that I can roughly organize my work toward the completion of this project.

- Prepare a proposal and survey important literature
- Dig deeper into the literature and search for implementational details
- Clean, organize and explore the data. Include or remove clinical covariates as necessary.
- Prepare the code, run and test it extensively on example data. Decide if simulations will be necessary.
- Prepare a second draft with a detailed introduction and some results based on exploration.
- Run the alogrithms on the wheat data set, and collect the results. Prepare summaries and visualizations.
- Prepare Final Report

4. REFERENCES

[1] Silke Szymczak, Joanna M. Biernacka, Heather J. Cordell, Oscar GonzÃąlez-Recio, Inke R. KÃűnig, Heping Zhang, and Yan V. Sun. Machine learning in genome-wide association studies. *Genetic Epidemiology*, 33(S1):S51–S57, 2009.

[2] Xiang Zhang, Shunping Huang, Zhaojun Zhang, and Wei Wang. Chapter 10: Mining Genome-Wide Genetic Markers. *PLOS Computational Biology*, 8(12):e1002828, December 2012.

[3] Gustavo de los Campos, John M. Hickey, Ricardo Pong-Wong, Hans D. Daetwyler, and Mario PL Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345, 2013.

[4] Valentina Moskvina and Karl Michael Schmidt. On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology*, 32(6):567–573, September 2008.

[5] Erin Austin, Wei Pan, and Xiaotong Shen. Penalized Regression and Risk Prediction in Genome-Wide Association Studies. *Statistical analysis and data mining*, 6(4), August 2013.

[6] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006.

[7] Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege, and Andrew Collins. Machine learning approaches for the discovery of geneâĂŞgene interactions in disease data. *Briefings in Bioinformatics*, 14(2):251–260, March 2013.