A Machine Learning Method for Selection of Genetic Variants and Prediction of Type 2 Diabetes Mellitus Using Next-Generation Sequencing Data

Jung, Luann (School: Manhattan High School)

Type 2 diabetes mellitus (T2DM) affects the lives of millions of people through its life-altering complications. Worldwide, 3.4 million people die of diabetes annually. It is known that T2DM results from a combination of both environmental and genetic factors, but efforts thus far in studying the disease's mechanism have not been able to fully understand how the two factors interplay. Additionally, a 2016 Nature Reviews article summarized that the accuracy of predicting future type 2 diabetes from genetic polymorphisms is very low (AUC >0.68) at the population level. As such, innumerable associations between genes, environmental factors, and T2DM remain to be discovered. This research presents a method to identify subtle effects of genetic variants using whole genome sequencing data and improve prediction accuracy of T2DM at the population level. To achieve this, a new feature selection procedure and a classifier were proposed. The method involves (1) first applying sparse principal component analysis (PCA) to genotype data to obtain orthogonal features; (2) using SNP-specific regularization parameters to reduce the false positive rate of feature selection; (3) verifying feature relevance through lasso penalized logistic regression in conjunction with sparse PCA. After application to a dataset containing 625,597 SNPs and 23 environmental variables from each of 3,326 humans, the method identified 466 genetic variants that have subtle effects on T2DM prediction. These variants, in conjunction with clinical characteristics, led to greatly improved prediction accuracy (AUC 0.79) for new patients at the population level. The proposed method also has the advantage of computational efficiency, and thus provides a promising tool for large-scale genome-wide association studies.

Awards Won:

Fourth Award of $500