Modeling the Risk Factors of Heart Disease

Kate Sanders and Robert Long

August 18, 2017

According to the American Heart Association, an estimated 16.5 million Americans over the age of 20 suffer from Coronary Heart Disease (CHD). Approximately 1 in 7 deaths that occur in the United States are a result of CHD, making it the country's leading cause of death. Because of this, many resources have gone into researching the causes of CHD. Traditional studies take years to conduct and only analyze the effects of a very small number of risk factors on CHD development. With the recent advances in Data Mining algorithms, it is now possible to model the connections between multiple complex risk factors of CHD.

Kate Sanders, a sophomore studying Computer Science at Hendrix College, and Robert Long, a senior studying Computer Science at DePauw University, spent 10 weeks using machine learning to model the risk factors of CHD. They worked with Nandini Ramanan and Dr. Sriraam Natarajan to create Bayesian Networks that modeled the relationship between 16 CHD risk factors and Coronary Artery Calcification (CAC), a primary indicator of CHD. This research was part of the Indiana University ProHealth Research Experience for Undergraduates, a program sponsored by the National Science Foundation.

The researchers used data from the CARDIA database, which was collected through a 20 year study involving over 5,000 participants. Research was focused on year 20 data, at which time all participants were aged 38 to 50 and approximately 11% had CAC. The researchers modeled 16 CHD risk factors and CAC using the R package bnlearn, which performed both network structure and parameter learning using several algorithms. An intersection of the Bayesian Networks created using 3 scoring metrics and the Hill-Climbing structure learning algorithm revealed several interesting correlation.

After learning the parameters of the intersection network, the researchers found that sex and glucose levels played the largest role in CAC development. Those with high glucose levels were more likely to develop CAC. Men where much more likely to have CAC than women. Glucose levels were primarily influenced by a person's BMI and sex, with overweight individuals more likely to have high glucose levels. The researchers also observed interesting correlations between a patient's race, education level, and whether they smoked.

When asked about the impact of this sort of research, Dr. Natarajan stated, "Our grand vision is to enable machine learning to empower people in taking control of their cardiovascular health." Future research will involve modeling

the changes in risk factors over time, as well as including more socioeconomic data in the modeling process.

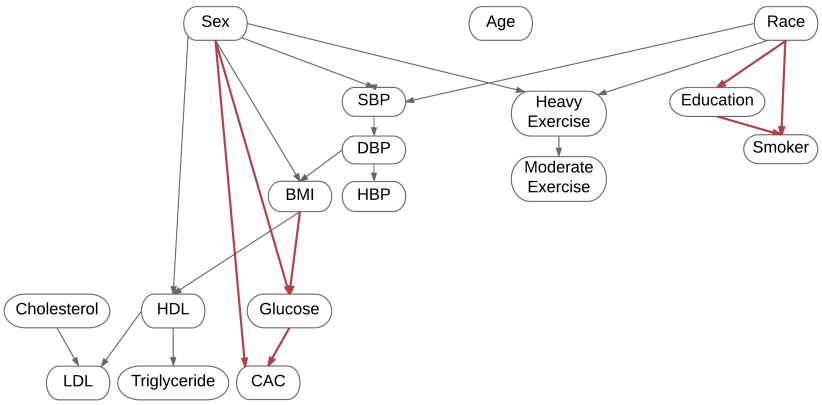

Figure 1: Intersection of models learned using the AIC, BIC, and BDe scoring metrics in the Hill-Climbing algorithm with year 20 CARDIA data.

This research was funded by the National Science Foundation Grant CNS-1560276 as part of the ProHealth REU at Indiana University. For more information, contact Kate Sanders at sanderskm@hendrix.edu or Robert Long at robmlong95@gmail.com.