Machine Learning Approaches for Prediction of Preterm Birth

Ilia Vovsha, Ashwath Rajan, Ansaf Salleb-Aouissi, Axinia Radeva, Hatim Diab, Ashish Tomar
Columbia University Center for Computational Learning Systems
Clinical Informatics Group (CING) cing@ccls.columbia.edu

Ronald Wapner
Department of Obstetrics and gynecology, Columbia University Medical Center

Anita Raja
University of North Carolina (UNC) Charlotte

Mary McCord
Medical College//Children's Hospital of Wisconsin

Tara Randis
Department of Neonatology, Columbia University Medical Center

The United States spends over 26 billion dollars per annum on the delivery and care of the 12-13% of infants who are born preterm (Berhman et al. 2007). As preterm birth is a major public health problem with profound implications on society, there would be extreme value in being able to identify women at risk of preterm birth during the course of their pregnancy.
Previous research has largely focused on individual risk factors correlated with preterm birth (e.g. prior preterm birth, race, and infection) and less on combining these factors in a way to understand the complex etiologies of preterm birth. Today, there is no widely tested prediction system that combines well-known factors (Davey et al., 2011) with a good prediction technique to provide actionable decisions within a clinical environment (Mercer et al., 1996).

Our ongoing research addresses this gap by conducting a deeper analysis of the preterm prediction study data collected by the NICHD Maternal Fetal Medicine Units (MFMU) Network, a high-quality data for over 3,000 singleton pregnancies having detailed study visits and biospecimen collection at 24, 26, 28 and 30 weeks gestation. Reports from this dataset used relatively straightforward biostatitistical methodologies such as relative risk assessments to measure associations between risk factors and PTB. These methods include descriptive statistics, Pearson correlation, Fisher's exact tests and linear/logistic regression where risk factors are studied independent of each other.

We conduced experiments on this MFMU dataset using non-linear Support Vector Machines to predict mothers at high risk of PTB at different time points, the main visits in the preterm prediction study. The generality of the models were assessed through cross-validation and then tested on a reserved subset that served as new (unseen) data. Our results demonstrate the superiority of non-linear methods in predicting preterm birth. Besides the fact that no time-dependent prediction was ever used on the MFMU dataset, We obtained an average of sensitivity and specificity in predicting PTB of 56% and 68% respectively, well above the ~21% for sensitivity and ~30% for specificity reported in the literature.

We also provide our initial efforts toward harnessing Electronic Health Records to prepare data for preterm birth prediction. We show our preliminary work on a 5-year snapshot of data for mothers and babies from the New York Presbyterian Hospital EHR systems. We give our initial data preparation and statistics w.r.t. preterm birth diagnostics. We stress the challenges faced in preparing EHR data for Machine Learning ranging from restoring the link between mothers and their babies to diagnostic validation.

We finally show how exciting is this application from a machine learning perspective ranging from multiple labels learning, temporal prediction, to multifactorial and privileged information, aspects we would like to pursue as a continuation of this effort.