Automatically Explaining Machine Learning Prediction Results: A Demonstration on Type 2 Diabetes Risk Prediction

Gang Luo
Department of Biomedical Informatics and Medical Education
University of Washington
luogang@uw.edu

Main Ideas

- A machine learning model achieving high accuracy is usually complex and gives no explanation of prediction results
- Challenge: Need to achieve high prediction accuracy as well as explain prediction results
- Key idea: Separate prediction and explanation by using two models concurrently
  - The first model makes predictions and targets maximizing accuracy
  - The second model is rule-based
    - Used to explain the first model's results rather than make predictions

Main Ideas – Cont.

- The rules used in the second model are mined directly from historical data
- Use one or more rules to explain the prediction result for a patient
- Suggest tailored interventions based on the reasons listed in the rules

Some Results

- Test case: Predicting type 2 diabetes diagnosis within the next year
- Electronic medical record data of 10K patients
- Can explain prediction results for 87% of patients who were correctly predicted by a champion machine learning model to have type 2 diabetes diagnosis within the next year

An Example Rule

- The patient had prescriptions of angiotensin-converting-enzyme (ACE) inhibitor in the past three years AND the patient's maximum body mass index recorded in the past three years is ≥35 → the patient will have type 2 diabetes diagnosis within the next year
  - ACE inhibitor is used mainly for treating hypertension and congestive heart failure
  - Obesity, hypertension, and congestive heart failure are known to correlate with type 2 diabetes
- Example intervention: Enroll the patient in a weight loss program