CSCI-UA 480: Predictive Analysis

Summer 2017, New York University

Professor: Anasse Bari

Predicting Early Stage of Chronic Kidney Disease Based on Blood Test Result

Author: Tom Li ([gl1147@nyu.edu)](mailto:gl1147@nyu.edu))

Project Description

---------------------------------------------------------------

What is CKD?

Chronic kidney disease (CKD) is a common disease that affects billions of people around the globe. It is a condition characterized by a gradual loss of kidney function over time. Kidney is one of the most important organs in human body. The major function of a kidney is to remove waste products and excess fluids from the body through urine. It is also responsible for other human body activities such as balancing salt levels as well as acids in the blood, producing hormone that stimulates red blood cell production, etc. In simple terms, kidney acts as a life-sustaining filter that eliminates wastes from our blood to keep it in a healthy state. Therefore, it is not hard to imagine if kidney fails and is left untreated, the consequences would be life-threatening. The first consequence of undetected CKD is the risk of developing progressive loss of kidney function that can lead to kidney failure (also called end-stage renal disease, ESRD) which means regular dialysis treatment or a kidney transplant is needed to survive. The second consequence of CKD is that it increases the risk of premature death from associated cardiovascular disease (i.e. heart attacks and strokes). Individuals who appear to be healthy who are then found to have CKD have an increased risk of dying prematurely from cardiovascular disease regardless of whether they ever develop kidney failure.

CKD can be caused by hypertension and diabetes. However, kidney disease starts slowly and silently and progresses over years. There are five stages of CKD and it is not easy to be detected in early stages because usually there is no signs or symptoms. People with CKD might not go through all 5 stages but it does not go away and can’t self-cure. Currently there is no cure for chronic kidney disease, although treatment can slow or halt the progression of the disease and can prevent other serious conditions developing.

How do we predict CKD?

As stated in the previous section, CKD can be dangerous when it gets to late stage. So it is very crucial for doctors to be able to give a concrete diagnose in early stage before the kidney starts to fail and the symptoms start to appear. This project aims to predict if a patient has CKD in early stage based on their blood test result. In order to achieve this task, we need to first analyze what are the potential differences between a CKD test positive patient and a CKD test negative patient. Going back the function and the cause of CKD, it is safe to assume that a CKD patient tends to have higher levels in bacteria, sugar, blood pressure, sodium, potassium and lower level of red blood cells in their blood. Thus a blood test would be extremely helpful in predicting CKD.

Fortunately we are provided with a dataset of 400 Indian patients (250 CKD test positive, 150 CKD test negative). This dataset can be found on the UCI Machine Learning Repository website or <https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease> and was created by Dr. P.Soundarapandian.M.D.,D.M in July 2015. The dataset has 25 attributes including a class attribute. Most of the attributes are included in a regular blood test report and the rest can be easily collected from a patient.

Cross Industry Standard Process for Data Mining (CRISP-DM) is applied to this project. The work that will be done in each phase is briefly explained in the following:

Phase one Business Understanding: CKD is a common disease that exists among people around the world and it has a potential that it will lead to kidney failure. It is very important for doctors to be able to diagnose CKD in its early stages and to prescribe proper drugs that treat this condition. If successful the model built in the project could profit many pharmaceutical companies and most importantly, it can save millions of lives.

Phase two Data Understanding: As stated above, the data source is UCI machine learning repository. The dataset has 400 instances (patients) and 25 attributes including a class attribute which indicates if a patient has CKD or not. It also has missing attribute values.

Phase three Data Preparation: Missing values will be filled with average and the data will be reduced by selecting a subset of its features that are most relevant to analysis.

Phase four Modeling: A classifier is built through classification algorithm (C4.5).

Phase five Evaluation: Classifier is tested for accuracy.

Phase six Deployment: Model will be used in real life predicting CKD and a full documentation will be included.

Tentative Timeline

* June 1-15: Business & Data Understanding;
* June 16-30: Data Preparation;
* July 1-31: Modeling;
* August 1-9: Deployment & Final Report

Team Member

* Tom Li (solely responsible for all tasks listed above)