My Data Science Challenging Problems # 55: Imbalanced Classification

Drafted by Haoda Fu

XXX

XXXX

xxx

May 8, 2020

Abstract

If you happen to have a solution of this question, please contact Haoda Fu at fu_haoda@lilly.com or fuhaoda@gmail.com or commit your solution to `https://github.com/fuhaoda/StreamingDataFeatureSelections.git`

Key words and Phrases: *** ***.
Short title: ***

1 Introduction

This problem is about how to solve a classification problem when the observed positive and negative cases are imbalanced.

Let us assume our samples are representative of the population at this moment. Suppose we have $\{(Y_i, X_i), i = 1, \cdots, n\}$ where $Y_i \in \{0, 1\}$ and X_i are covariates. Data can be split into training and testing datasets, and we use $\mathcal{M} = \{Y_m, X_m\}$ for training (m stands for modeling) and use $\mathcal{H} = \{Y_h, X_h\}$ as testing (h stands for holdout).

As we know, the misclassification in reality often has different cost. How can we develop a classifier for a given cost for

- NPV vs PPV
- Sensitivity vs Specificity

To be specific, how we can estimate $D_o(\cdot)$ as a classifier built on $\mathcal{M}$ to minimize the loss calculated on $\mathcal{H}$ as,

$$L_o = \sum_{\mathcal{H}, Y_i=0} W_p I\{D_o(X_i) \neq Y_i\} / \sum_{\mathcal{H}} I\{D_0(X_i) = 1\} + \sum_{\mathcal{H}, Y_i=1} W_n I\{D_o(X_i) \neq Y_i\} / \sum_{\mathcal{H}} I\{D_0(X_i) = 0\},$$

where W_p is the loss of false positive and W_n is the loss of false negative. The L_o can be calculated by multiple random split then taking the average loss as in cross validation.

Similarly, we can also ask for a classifier $D_t(\cdot)$ to optimize the weighted cost based on sensitivity and specificity calculated on $\mathcal{H}$.

To derive $D_0(\cdot)$ or $D_t(\cdot)$, we can have two approaches. One approach is based on the modified loss function and find some convex approximation. Another approach might be data augmentation approach.

For the data augmentation approach, what is a procedure that can be generically used for any existing classification algorithms, such as XGBoost, Deep Neural Network, Random Forest?

2 Sampling Bias

What if we know that positive or negative cases are under sampled with a know ratio $r(X)$. How we can take this into account for estimating the classifier.

3 Personalized Solution

What if we want to get an optimal solution for a fixed X.

4 Personalized Variable Selection

To achieve certain level minimal accuracy, we may not need to measure all the variables. How we can measure right amount of number variables to minimize the measurement cost. The solution is likely a sequential variable selection solution.

Appendix: Proof of Equation ()

.