Supervised Interactive Learning on Biomedical Documents Classification

Yu Zhang
Department of Computer Science
University of Kentucky

The Biomedical Problem

Cancer abstract:

"Extended surgery and pelvic exenteration for locally advanced rectal cancer. What are the limits? Historically, locally advanced rectal cancers with invasion of tumor into adjacent organs (T4 N1, 2 tumors) have been considered poor prognosis cancers treated with palliative intent. However with the advent of multi-modality therapy and improvement in surgical reconstructive techniques, extended resections for rectal tumors are possible with acceptable patient morbidity and excellent oncological outcomes."

Non-Cancer abstract:

Pre-employment integrity testing across multiple industries. Despite the robust meta-analytic data available, very little comparative research exists on validities of integrity measures within specific industries. Among a sample of 2456 Israeli job applicants, integrity scores were found to be significantly correlated with self-reported counterproductive work behaviors across eight different industries, with no evidence of adverse impact by gender, age, or national origin. These results are believed to be of practical importance to the diverse organizations administering integrity tests.",nocancer

The Biomedical Problem

Cancer abstract:

"Extended surgery and pelvic exenteration for locally advanced rectal cancer. What are the limits? Historically, locally advanced rectal cancers with invasion of tumor into adjacent organs (T4 N1, 2 tumors) have been considered poor prognosis cancers treated with palliative intent. However with the advent of multi-modality therapy and improvement in surgical reconstructive

We want to do binary classification on abstracts.

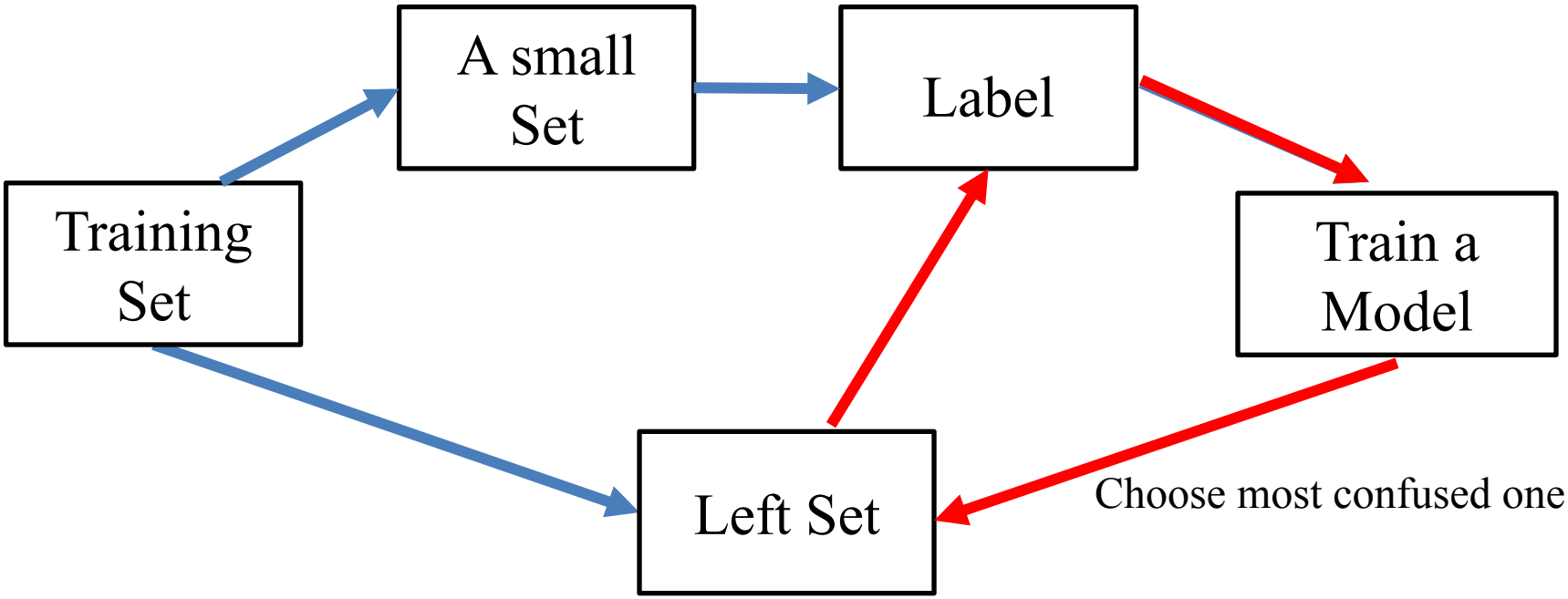
Pre-employment integrity testing across multiple industries. Despite the robust [illegible] analytic data available, very little comparative research exists on validities of integrity measures within specific industries. Among a sample of 2456 Israeli job applicants, integrity scores were found to be significantly correlated with self-reported counterproductive work behaviors across eight different industries, with no evidence of adverse impact by gender, age, or national origin. These results are believed to be of practical importance to the diverse organizations administering integrity tests.",nocancer

Supervised Learning

Non Interactive Machine Learning:

Supervised Interactive Learning

Interactive Supervised Machine Learning:



Why Supervised Interactive Learning

Save human label work!!!

Human only label the sample which machine is most confused on, so the machine will learn faster and better.

Limited labeled samples will be enough for machine learning.

Contribution

1. Using the interactive machine learning method to reduce the expert label work on biomedical documents.

2. The interactive part is added to two different supervised method as Naïve Bayes and SVM.

3. Write the source code of Naïve Bayes which can be trained online.

Evaluation

Data Set
Training set:1330 biomedical paper abstracts.
Testing set:200 biomedical paper abstracts.
Each one is labeled as “cancer” or “nocancer”.

The word “cancer”, “Cancer”, “cancers” are removed in the training set. We want to increase the challenge to our method In the evaluation.

Our metric is precision, recall and F-score.

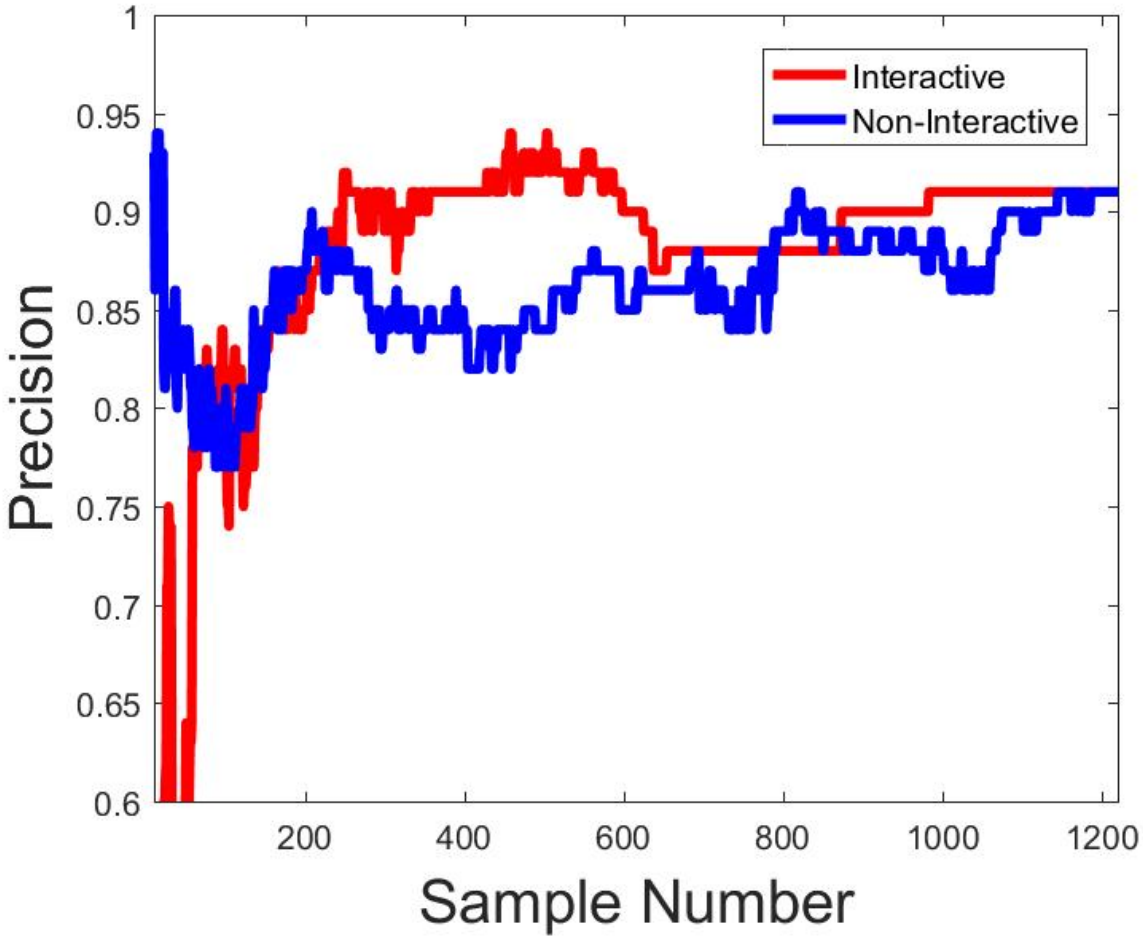
(The dataset used in this project is from the course of BMI 733, University of Kentucky.)

SVM Interactive Learning

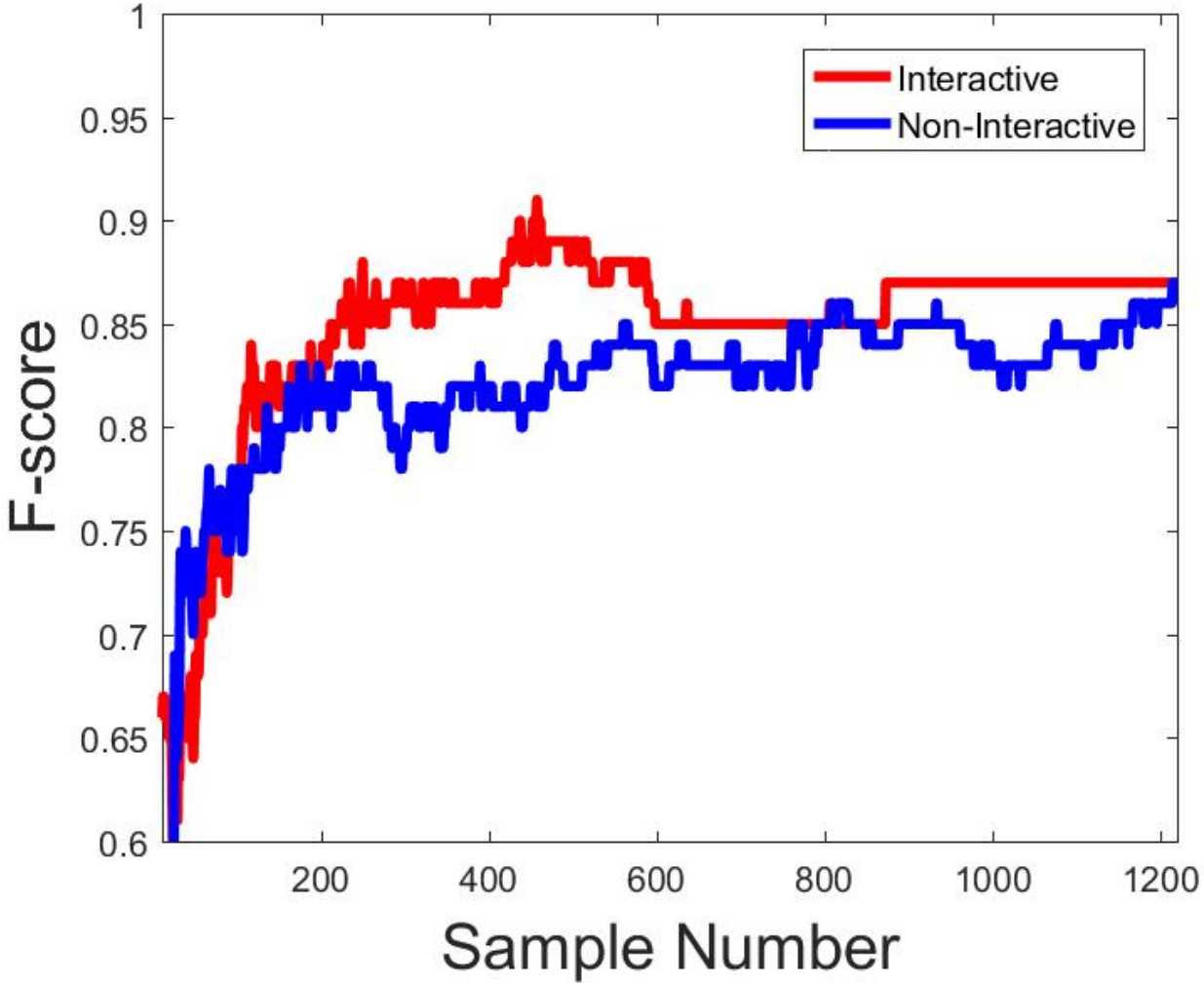
Interactive Learning:
Repeat{
    Choose the most confused sample *min(|proba-0.5|)*
    Retrain SVM model.
    Gives the P/R/F-score on testing set.
}

The SVM is implemented by sk-learn API.
The kernel is “linear”.

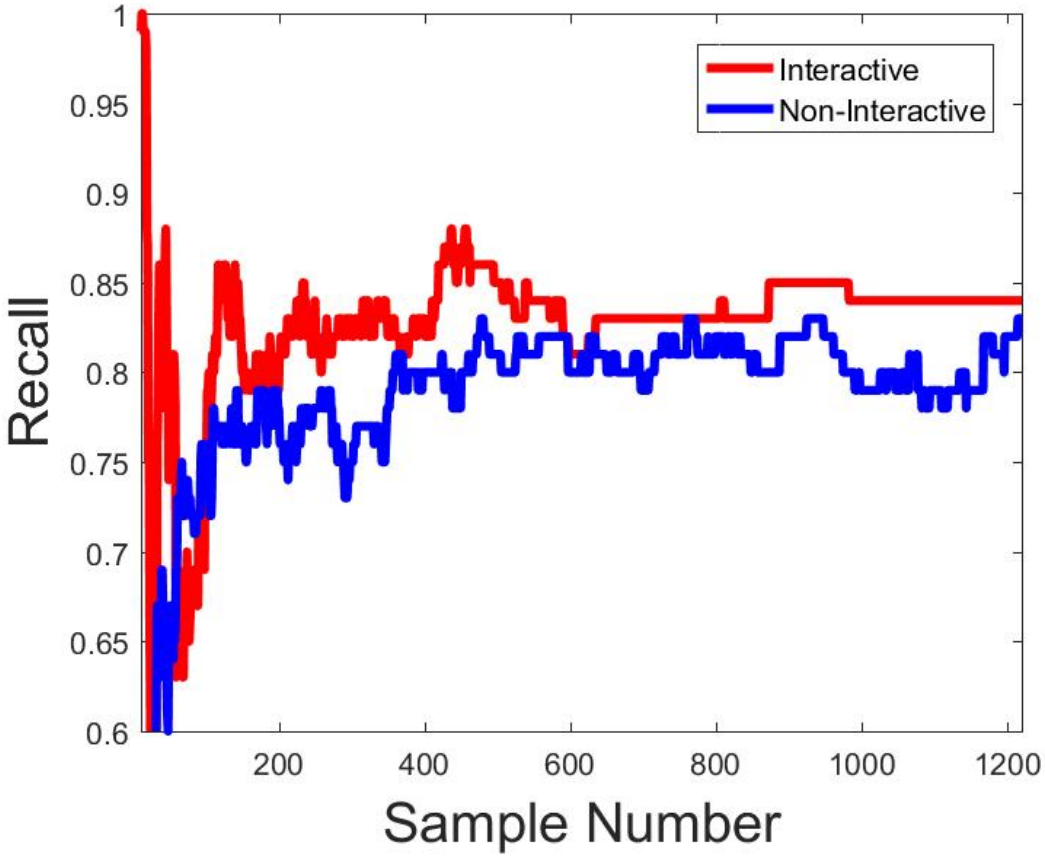
SVM Result


Interactive
Non-Interactive
Precision
Sample Number
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
200
400
600
800
1000
1200

SVM Result


Interactive
Non-Interactive
F-score
Sample Number
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
200
400
600
800
1000
1200

SVM Result


Interactive
Non-Interactive
Recall
Sample Number
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
200
400
600
800
1000
1200

Thanks!

- Question and Answer