Cervical cancer key risk factors identification

Capstone_Project_Proposal

Machine Learning Engineer Nanodegree
Tim
January 30th, 2018

Domain Background

Cervical cancer caused by uncontrolled growth of cells in the cervix (Reference), its' death rate might be higher than people previously thought it to be (Reference). In the United States, an estimated number of death of this disease in 2018 for women is up to 4170 (Reference). So it's important to find out which key risk factors would signally increase women's chance of developing cervical cancer, as well as to predict whether someone got the disease. So that we can come up with better preventive methods(and diagnostic) to detect the disease earlier, and minimize the death rate of it.

Problem Statement

The data set provides historic medical records (described by few attributes) of 858 patients (The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela). The goal is first to use the data to train a classfier that can be used to tell whether a patient has this disease when new data flow in, and to tell among all these features, which ones are going to be more predictive (or maybe say informative) than others in diagnosing cervical cancer.

Since the data set is a labeled one (each data point has been labeled 1 for having the disease, 0 for not having the disease), so the problem is binary classification in supervisd learning. A few classification algorithms should be explored in the purpose of finding the most suitable one.

Datasets and Inputs

The data set to be used in this project has been downloaded from UCI Machine Learning Repository. Information of those features can be found above link. I would also explain some terminology which have been used to name those features.

- Intrauterine Device(IUD) is a device that to be placed into a woman's uterus to prevent pregnancy, detailed description can be found through the wiki page for this term IUD
- Sexually Transmitted Disease(STD) is disease that are spread by sex. STD
- Diagnosis(Dx).Dx
- Hinselmann refers to the medical diagnostic procedure which has been developed by the German physician Hans Hinselmann, its' main application is to prevent cervical cancer. More details can be found in Colposcopy
- Cervical intraepithelial neoplasia(CIN) is an indicaition of cancerization, more details can be found here CIN

- Schiller refers to a medical test to diagnose cervical cancer, but it's not specific for cervical cancer. Schiller
- Cytology and Biopsy are both tests that industry use to find cancer, more details are available here Cytology and Biopsy

This data set contains 36 relevant features with up to 858 instances, it has demographic, habits (like smoking), and medical records which would be useful for prediction. What's more, there're already some kernels on Kaggle working on this subject so we can use them as benchmark. In consideration of these, the data set should be an appropriate one for the project.

Citation:

Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.

Solution Statement

We've labeled data set and the task associated with it is a binary classification one, some supervised learning algorithms should be approprate approaches to tackle this problem, more precisely, I would like to try things like SVM, Gradient Boosting, or GaussianNB. Accuracy score and F-score would be our quantifiable result which can be used to compared with some benchmark models.

Benchmark Model

The data set has been used to set up a competition on Kaggle, many kernels have achieved an accuracy of 95% (more or less). So I would like to use 95% accuracy as the reference to evaluate the performance of my model, its' accuracy should near around 95%, but it should not lower than 93%.

Evaluation Metrics

As I mentioned, I would use accuracy score to evaluate the performance of both benchmark and solution model. What's more, in consideration of the domain background, I believe it's more important for our solution model to correctly identify those patients that actually have the disease than to recall them, so I would also use F-score as evaluation metric as well.

Project Design

1. Data Exploration - At this stage I'll try to have a preliminary understanding of the data. E.g. total number of records, how many positive and negative cases in the data, and whether some features contain way too many missing values that might be worthless to our study and better to be dropped off.
2. Data Preparation - Data cleansing will be conducted, missing value should be taken care of (maybe to drop off some of those insignificant data points and featrurs), and

might need to transform some skewed continuous features to normalized the data range, and to split the data into training and test set of course.

3. Model Performance Evaluation - In this part I'll try on few algorithms and determine which one is more suitable at modeling the data by comparing their accuracy score and F-score on test set, and employ k-fold here due to the limited amount of data.
4. Improving Results - At this final stage, I'll choose the best fit model from all the candidates at last stage, a grid search optimization might be employed for target model sourcing as well as model tuning (parameters)