Term Project Proposal:Applied Machine Learning Algorithm For Breast Cancer Prediction

Qingyang Li
School of computing and information,
University of Pittsburgh
Pittsburgh, PA 15213
Email:qil77@pitt.edu

***Abstract*—As the requirements of the term projects, one of the four topic should be selected as our final project. Based on the machine learning course materials I learnt this semester so far, I am interested in applied machine learning algorithms on real-life data, make the predictions and evaluate their performances. Based on this motivation, after doing some searching and consideration, I found a publicly available data set that contains images of breast cancers specimens. I decided to use this data set to train multiple machine learning algorithms, evaluate their performance and choose one has the best performance.**

I. Introduction

A. Background

Since I was a child, I often heard a lot from my parents that their friends or some people they knew were suffered from breast cancer. In fact, according to CDC, breast cancer is the second most common cancer among women in the United States[2], and the major cause of death in women worldwide. However, the number of women who have died of breast cancer has decreased by 42% from 1989 to 2019 thanks to early detection and treatment improvements.[1] It is non-trivial that apply machine learning to medical records can be an effective tool to predict the probability of a person who could have a breast cancer in the near future. And help doctors to give the early intervention before the breast cancer turn to the late stage.

B. Breast Histopathology Images

The data I choose is from Kaggle.[4]. The data set consisted of 277,524 patches of size 50x50 images that were classified as invasive ductal carcinoma (IDC) positive and negative (198,738 IDC negative and 78,786 IDC positive). Invasive ductal carcinoma is the most common subtype of all breast cancers.[3]. Loading the whole data set is really cost, due to the limitation of my personal computer, I decreased the sample to 10,000 patches with 7160 IDC negative and 2840 IDC positive at the moment until I have more available computing resources to use.

C. Machine Learning Classification

The most methods that used to solve binary classification problem are logistic Regression, k-nearest neighbors, decision trees, support vector machine. And convolutional neural network (CNN) is most commonly applied to analyze visual imagery in deep learning. To experiment with advanced ML methods, those features are applied to basis machine learning algorithms as following:

- K Nearest Neighbor Search on a KD tree: KNN intend to search the nearest neighbors for a target in the entire training set, therefore the prediction can be time consuming. KD-trees are a specific data structure for efficiently representing the training data. In particular, KD-trees helps partition the data points based on specific conditions to reduce the time.
- Random Forest: Random forest is a tree-based machine learning algorithm that uses multiple decision trees for making decisions, each node in the decision tree works on a random subset of features to calculate the output. Those outputs then be combined to generate the final output.
- SVM with Gaussian kernel: Applying kernel function can transform the non-linear decision surface to a leaner equation in a higher number of dimension spaces. Gaussian kernel are universal kernels, their use with appropriate regularization that optimal the predictor which minimizes both the estimation and approximation errors of classifier.
- CNN with AlexNet architecture. AlexNet is a famous CNN architecture that was the first major CNN model used GPU's for training. It consisted of 8 layers and used ReLu activation function.

The above advanced algorithms are completing with logistic regression for the best performance in this data set.

II. Related Work

My experiment is inspired by a survey about involvement of machine learning for breast cancer image classification [5]. The survey listed a series of classification methods can be applied to the breast cancer prediction,different features selection and some references of other experiments and results that used different CNN methods to predict breast cancers on different data set.

III. Experiment Setting

I plan to select logistic regression as the baseline model since it's the simplest classification model in supervised learning algorithms. Then I will compare the results from baseline model with other models(KNN-KD, Random Forest, SVM with gaussian kernel,AlexNet) using confusion matrix

measuring precision,recall and F-measure, ROC area under the curve and precision-recall (PR) area under the curve as evaluation methods to show the model's performance. The python libraries for model implementation for the experiment are mainly scikit-learn and tensorflow.

IV. Schedule of Milestones

A. March 21-March 27

This week I will do image process and understand the potential issues in these models. Read paper about related experiments. Try to increase the sample sizes to test the limitation about my computer. Wait for the proposal feedback and make adjustments to the experiment settings.

B. March 28-April 3

Code implementation for the models.

C. April 4-April 10

Training the models and adjust the models based on the training results.

D. April 11-April 17

Testing the models and adjust the models based on the testing results.

E. April 18-April 24

Final wrap up, results analysis and report preparation.

References

[1] *Breast cancer statistics by Cancer.net*. Feb. 2022. URL: https://www.cancer.net/cancer-types/breast-cancer/statistics.

[2] *Breast cancer statistics by CDC*. June 2021. URL: https://www.cdc.gov/cancer/breast/statistics/index.htm.

[3] *Invasive Ductal Carcinoma*. URL: https://www.breastcancer.org/types/invasive-ductal-carcinoma.

[4] Paul Mooney. *Breast Histopathology Images*. 2017. URL: https://www.kaggle.com/paultimothymooney/breast-histopathology-images.

[5] Kong Y Nahid AA. *Involvement of Machine Learning for Breast Cancer Image Classification: A Survey*. 2017. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5804413/.