Machine Learning for Breast Cancer Diagnosis

Capstone Project – Machine Learning Engineer Nanodegree

Author: Harini Sridhar

Domain Background: Breast cancer, one of the leading causes of cancer deaths in women today, is a condition in which malignant (cancer) cells form in the tissues of the breast. In 2018, an estimated 266,120 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S and about 40,920 women are expected to die [1]. The death rates have been dropping steadily since 1989, as a result of treatment advances, earlier detection through screening, and increased awareness.

The most common breast cancer screening test is a mammogram, which is an x-ray of the breast. The ability of a mammogram to find breast cancer may depend on the size of the tumor, the density of the breast tissue, and the skill of the radiologist. However, scanning through mammograms for diagnosis is time-consuming and challenging even for expert radiologists. Findings on a mammogram leading to further recall are identified in approximately 5%-10% of patients, even though breast cancer is ultimately confirmed in only three to ten cases in every 1000 women screened. To overcome the known limitations of human observers, additional reading of mammograms by another radiologist are often obtained before making the final diagnosis. However, this approach is neither scalable nor financially viable. Computer technology is proving to be helpful in this regard. One very promising adaptation is computer-aided detection (CAD) in mammography [2]. A CAD system is usually comprised of the following stages:

- *Candidate generation* – identifies suspicious unhealthy candidate regions of interest (ROI) from an x-ray image;
- *Feature extraction* – computes descriptive features for each candidate so that they can be represented by a vector of numerical values or attributes;
- *Classification* – differentiates candidates that are malignant cancers from the rest of the candidates based on the extracted features; and
- *Visualization* – presentation of CAD findings to the radiologist.

Problem Statement: The goal of this project is to utilize machine learning to predict if a *candidate* tumor region from the x-ray is malignant or benign. For this purpose, we use a dataset that consists of x-ray images from patients with a benign or malignant tumor. More specifically, the ROI definition and feature extraction stages were carried out *a priori* using standard tools, and this project will build a predictive modeling solution for diagnosis based on the features.

This dataset presents crucial challenges that need to be considered during the solution design: First, the dataset has an imbalanced class distribution. Only a very small fraction of the dataset is actually malignant. Most popular classification methods are effective only when the datasets are balanced. Consequently, it is important to understand their behavior in this constrained scenario. Second, the large number of features in the data can make model design challenging, in terms of generalizability. Third, the dataset has some inherent challenges (multiple scans for the same patient), thus requiring extensive pre-processing. The project aims to overcome these challenges and build a good predictive model.

Datasets and Inputs: The dataset used for this project was curated as part of the KDD Cup competition organized in 2008 [3] and was originally provided by Siemens Medical Solutions. A breast cancer screen typically consists of 4 X-ray images; 2 images of each breast from different directions giving a total of 4 images per patient. Each image is represented by several candidates. Each candidate, has 117 features, and a class label indicating whether or not it is malignant. The dataset consists of a total of 102,294 candidates corresponding to 1712 patients, but only an extremely small fraction of these candidates (~650) is actually malignant. Following common practice in machine learning, this dataset will be split into training and validation sets for analysis and benchmarking. The dataset can be downloaded from [4].

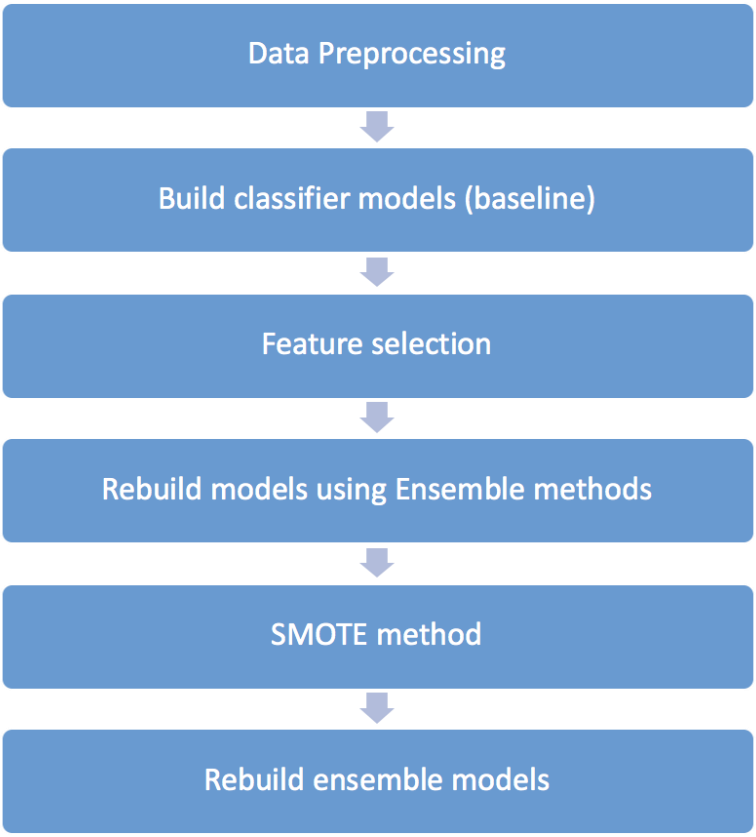
Solution Statement: The data will first be preprocessed and split into train/test sets. Different classifier models will be fit, along with hyper-parameter tuning, to evaluate their effectiveness on imbalanced datasets. In order to achieve variance reduction, we will design an ensemble classifier, wherein the best performing classifiers from the previous step will be used as the base model. The SMOTE method will be used to deal with the dataset imbalance and the ensemble classifier models will be rebuilt. The designed models will be evaluated using metrics such as precision/recall, and free response receiver operating curves (FROC).

Benchmark Model: Different classifier models such as decision trees, linear SVM, kernel SVM, logistic regression and nearest neighbors will be fit to the data. Hyper-parameters will be tuned and the results will be used to set up a strong baseline. Evaluation metrics such as precision and recall, and FROC will be used.

Evaluation Metrics: There are several metrics that can be considered when we evaluating the quality of predictive models.

- Accuracy: Overall accuracy is a bad metric for imbalanced datasets, since it is overly optimistic. For example, the overall accuracy can be over 99% without detecting even a single malignant case. Hence, this metric will not be used in this project.
- Precision and Recall: The precision and recall of individual classes can be chosen as a metric. In medical diagnosis scenarios, false negatives are more severe than false positives. Hence, we want to achieve high recalls.
- FROC: An ROC curve indicates the true positive rate (sensitivity) as a function of the false positive rate (1-specificity). FROC is similar to ROC, except that the false positive rate on the x-axis is replaced by the number of false positives per image. Finally, we measure the area under the curve.

Project Design



The proposed project will involve the following steps:

1. **Data preprocessing**: The dataset is known to have missing values. These will be addressed in this step. If there are outliers, they will be addressed too. The data will be split into training and testing sets.

2. **Build classifier models**: Different classifier models such as decision trees, linear SVM, kernel SVM, logistic regression and nearest neighbor will be fit to the data. Hyper-parameters will be tuned and this will be used to identify strong baselines.
3. **Feature selection**: Since there are a large number of features, feature selection will be used to reduce the dimensionality and improvements, if any, will be studied.
4. **Rebuild models using Ensemble methods**: Ensemble methods reduce the variance of complex classifiers so that overfitting will not occur. Methods like bagging classifiers, AdaBoost and XGBoost (only on decision trees) will be used on the best (maybe top 2) of models from Step 3 to improve classification.
5. **SMOTE method**: Since the data is highly imbalanced, sampling based ideas will be ideal. There are two kinds of sampling that can be used. One is by undersampling the negative samples. Another method is the SMOTE method [5]. Here, the negative samples are undersampled and the positive (malignant) samples are oversampled.
6. **Rebuild ensemble models**: The ensemble models are rebuilt and the impact of using the SMOTE method will be studied. Inferences from this project will be applicable to a wide range of problems that are characterized by complex classification boundaries and highly imbalanced class distributions.

References

[1] http://www.breastcancer.org/symptoms/understand_bc/statistics
[2] http://www.kdd.org/kdd-cup/view/kdd-cup-2008/Tasks
[3] http://www.kdd.org/kdd-cup/view/kdd-cup-2008/Data
[4] http://www.kdd.org/cupfiles/KDDCupData/2008/training_data.zip
[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. JAIR 16 (2002), 321–357