Machine Learning Capstone Project Proposal

Domain Background

Breast cancer is the second leading cause of cancer death among women, only behind skin cancer. Every two minutes, a woman is diagnosed with breast cancer and every thirteen minutes a woman will die from it. It is estimated that 1 in 8 women in the U.S. will be diagnosed with breast cancer in their lifetime, this translates to 252,710 diagnoses every year in the U.S. alone [1].

Early screening and detection are crucial to survival. If breast cancer is detected early, there are more treatment options resulting in a 93% higher survival rate in the next five years [2]. The size of a breast cancer and how far it has spread are key factors in predicting the prognosis [2].

Problem

Through medical imaging techniques such as mammograms, breast cancer can be detected visually by doctors before they can be felt in self-exams. However, there are not enough doctors to examine the results thoroughly due to the sheer number of examinations. With the shortage of doctors expected to last through 2030 [3], it becomes increasingly important for women to get access to fast and reliable screenings for breast cancer.

Machine learning techniques can be utilized to analyze images of breast mass obtained through various methods to detect the presence of tumors and more importantly, whether the tumor is benign or malignant. Benign tumors do not spread throughout the body and requires different treatment as opposed to malignant tumors, which is cancerous and has the ability to spread uncontrollably to surrounding issue [4]. This can be used to assist doctors by acting as a pre-screening check before they analyze the image themselves for confirmation, saving them time and acting as a sanity check. Doctors gain their expertise through years of training and experience, examining hundreds and thousands of images of breast cells to determine whether breast cancer exists based on the size, shape, smoothness, and other characteristics.

Dataset

The Diagnostic Wisconsin Breast Cancer Database is a dataset compiled by members of the University of Wisconsin General Surgery and Computer Science departments that has been made publicly available since 1995 [4]. The dataset contains 569 samples of features extracted from digitized images of fine needle aspirate of breast mass, which is a type of biopsy in which a thin needle is inserted into an area of abnormal tissue or body fluid. Each sample contains 32 features that describe the data. The dataset contains 357 samples of benign tumors and 212 samples of malignant tumors so it is not an even distribution, almost 63% of the samples are benign.

The features describe various characteristics of the cell nuclei present in the image along with the final diagnosis, whether the cell is benign or malignant. The features include measurements such as radius, texture, perimeter, area, smoothness, compactness, symmetry, etc. These

features are what doctors look for visually to determine whether a breast cancer cell is benign or malignant.

Solution Statement

The solution I propose is to utilize machine learning techniques to train a model on the dataset that can reliably and accurately predict whether a breast cell is benign or malignant (binary classification) based on the physical characteristics of the cell such as size, shape, smoothness etc. The model can be used as an automated pre-check before a doctor examines the image themselves and act as a second opinion for them. Metrics such as accuracy, sensitivity, specificity, and F-score of the model can be used to measure the performance of the model.

Benchmark Model

The benchmark model I will compare my solution against is the kaggle notebook by [6], which actually evaluates the performance of 3 different models on the same dataset. The 3 models investigated by [6] are:

1 Logistic regression
2 Decision tree
3 Random forest

[6] split the dataset into a 70:30 training/testing set. For the logistic regression and decision tree models, a select set of features were manually selected from the list of total features by manual inspection of the data and used to train the model. For the random forest model, [6] explored training using all the features, 5 top features, and a single feature.

The best model was shown to using random forest with these top 5 features:

1. concave points_mean
2. area_mean
3. radius_mea
4. perimeter_mean
5. concavity_mean

The model achieved a prediction accuracy of ~95% and a cross validation score of ~93%.

Evaluation Metrics

For this specific problem, it is important that we aim to maximize the true positives and negatives while minimizing the false positives and negatives. In other words, we'd like to predict that only breast tumors that are actually malignant as malignant, and the same for breast tumors that are benign. This can be measured using the F1-score, which is the harmonic average of the precision and recall and given by the equation:

$$2 * \frac{precision * recall}{precision + recall}$$

Since the distribution of the 2 classes in the dataset is skewed towards one class, the F1-score will also be the main metric used to measure the performance of the solution.

The final prediction accuracy on the testing set will be used as a supporting metric to evaluate our model's performance. The prediction accuracy will show how often our model is correct in predicting whether a tumor is benign or malignant. The equation to calculate accuracy is given by the equation:

$$\text{Accuracy} = (TP+TN)/\text{Total}$$

Where TP=# of true positives and TN=# of true negatives.

Project Design

The first step to any machine learning problem is to explore and try to understand the dataset to gain insight into any existing patterns. I will look at metrics such as mean, median, and mode for all the features, then look at them separated by class. I will try to see whether how the data is distributed across the features and classes by visually plotting them. It would also be a good idea to check for any missing data from the samples.

Next, I'll perform feature scaling for all of the features as part of the data pre-processing. Then, the dataset will be split into training and testing sets with either 70:30 train/test ratio, the same used by [6], or 80:20 train/test ratio. Since the dataset only contains 569 samples we may need more training data to prevent underfitting. I'll also explore using K-fold cross validation.

The dataset would be a good candidate for supervised learning techniques since the target variable (Diagnosis) is available. One of the supervised methods not covered by [6] is SVMs. I will train a SVM model on the dataset trying out different kernels and hyper-parameters and compare the performance. It would be good to evaluate the performance of the model when trained with all of the features versus a select set of features that are more important than the others, we can train the model on the top 1, 3, 5, and 10 features out of the total 32 features and see where the diminishing returns start to show. We can use the analysis from [6] to select the top important features.

For the SVM hyper-parameters gamma and C, I plan to do an exhaustive grid search to find the optimum values. A model will be trained using the optimum values of gamma and C found using grid search and the performance of the model will be measured through the evaluation metrics discussed in the previous section and compared with the results from [6].

Bibliography

[1] National Breast Cancer Foundation, "Breast Cancer Facts," National Breast Cancer Foundation, 2016. [Online]. Available: http://www.nationalbreastcancer.org/breast-cancer-facts. [Accessed 29 11 2017].

[2] Carol Milgard Breast Center, "Early Detection is Key," 2017. [Online]. Available: http://www.carolmilgardbreastcenter.org/early-detection. [Accessed 30 11 2017].

[3] S. Mann, "New Research Shows Shortage of More than 100,000 Doctors by 2030," 14 3 2017. [Online]. Available: https://news.aamc.org/medical-education/article/new-aamc-research-reaffirms-looming-physician-shor/. [Accessed 1 12 2017].

[4] T. Bollinger, "Benign and Malignant Tumors: What is the Difference?," 2017. [Online]. Available: https://thetruthaboutcancer.com/benign-malignant-tumors-difference/. [Accessed 4 12 2017].

[5] UC Irvine Machine Learning Repository, "Breast Cancer Wisconsin (Diagnostic) Data Set," 2007. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. [Accessed 1 12 2017].

[6] B. Waidyawansa, "Using the Wisconsin breast cancer diagnostic data set for predictive analysis," 3 12 2016. [Online]. Available: https://www.kaggle.com/buddhiniw/breast-cancer-prediction/notebook. [Accessed 5 12 2017].