Capstone Project Proposal

November 22, 2017

1 INTRODUCTION

1.1 Domain background

A cancerous tumour is a group of cancer cells that grow and destroy nearby tissues. Breast cancer is the type that usually starts in the cells of the breast. Over time, cells within the breasts may undergo changes and cease to grow or behave normally. These lead to non-cancerous (benign) breast conditions such as atypical hyperplasia and cysts, intraductal papillomas etc. But in some cases, these changes lead to breast cancer. Breast cancer usually starts in cells lining the ducts that carry milk from the glands to the nipple or in the cells of lobules, which are the groups of glands that make milk.

Breast cancer diagnosis usually begins with physical checks for the presence of lumps by feel and touch or through mammography. When this preliminary diagnosis signals any abnormailty, a biopsy is ordered to determine whether the tumor is benign or malignant. Clinical statistics indicate that only a small percentage of women with breast lumps actually get diagnosed with cancer.

Biopsy can be performed by several techniques. The most well-established technique among them is the Fine Needle Aspiration (FNA) Cytology. A FNA uses a 10 ml or 20 ml syringe and a 23 gauge to 27 gauge needle. The needle passes through the skin and into a palpable lump or area of breast abnormality and extracts the suspicious cells which are then assessed under the microscope by a pathologist. The technique comes with many advantages, it is easy to perform, is quick and inexpensive, as compared to an elaborate surgical biopsy. However, there are many disadvantages to it as well. The test does not distinguish between in-situ and invasive cancer, there is a significant rate of non-diagnostic samples but most importantly, the test is often associated with a significant `false negative rate`, especially when conducted by pathologists who are not so experienced.

While many of the disadvantages listed above are inherent to the nature of the FNA test, human errors resulting in false negative rates is definitely one aspect that will benefit from the application of machine learning techniques. Infact, since the tests result in a large number of features, it is quite necessary to have a develop a robust prediction model that can assess the test results and generate the result on whether the tumor is benign or not. Several researchers have attempted this and their work can be found in the References: [1-7]. Most of these researchers have tested their models on the widely used Wisconsin breast cancer dataset, which I will be using here as well. The dataset is available on the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/

1.2 Problem Statement

The problem in hand is thus a binary classification problem involving a large number of features. Here the inputs are the characteristics of the test culture (such as the radius, smoothness, compactness etc. of the distribution of cells), while the output is binary, i.e., benign or malign.

So the method to approach this would be to understand the significance of the features, execute some strategies for feature reduction, apply a binary classification algorithm and iterate this process, until performance saturates.

In short, the objective of this study is to build a predictive model that wil improve the accuracy, objectivity and reproducibility of breast cancer diagnosis by FNA.

1.3 Data Sets and Inputs

The dataset used in this work is the widely used Wisconsin University breast cancer dataset.

Features are computed from digitized images of fine needle aspirates (FNA) of breast masses. Specifically, the characteristics of the cell nuclei present in the images are recorded.

The database is available on Kaggle, as well as through the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/

It is also found on UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

The dataset has 569 rows and 33 columns. Amongst the 33 columns, the first two are `ID number` and `Diagnonsis (M=malignant, B = benign)`. And the last column is an unnamed column with only NaN values, so it is removed right away.

The other 30 columns correspond to mean, standard deviation and the largest values (points on the tails) of the distributions of the following 10 features computed for the cellnuclei;

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

All feature values are recoded with four significant digits.

The class distribution of the samples is such that 357 are benign and 212 are malignant, which is quite balanced.

1.4 Solution Statement

Since this is a binary-classification problem, it might be worth testing out - a simple Gaussian Naive Bayes approach.

But I am of the belief that plane separation approaches such as - SVM or logistic regression would be more relevant.

That said, it could also be a good idea to test out - ensemble methods such as bagging, random forests etc., - Non-parameteric cluster approaches as we might arrive at a relatively simple model with high performances.

Finally, - deep learning methods such as multilayer perceptron classifier with the Keras package could be tried out as well.

However, before applying any of the approaches above, a detailed analysis of the features have to be made. The following steps would be a good base model for the data exploration and analysis

- Feature transformation - standardizing,logarithmic transformation etc.
- Violin/Swarm plots to get a visual idea of which features are important and which ones don't matter much at all
- Joint plots of a few select features - this might help in the decision of how suitable a Gaussian naive Bayes model would be
- Feature selection - by the use of PCA or LDA, whichever might be more suitable
- Detection of outliers and the effects of keeping/removing them

1.5 Benchmark Model

As mentioned before, the Wisconsin breast cancer dataset is widely used and many researchers have applied machine learning techniques to determine the nature of a patient's lesion, whether malignant or not. A group in Turkey employed an SVM approach and reported higher than 99% accuracy, athough they used features such as single cell epithelial size, which is not usually included in a standard clinical report.~[1] Thus, there is still much interest and clinical relevance in the building of predictive models of this type.

Certain other models attempted by researchers include:

- the use of a neural network support Vector Machine (SVM) classifier with the Radial Basis Function (RBF) kernel and a fuzzy classifier
- combination of Decision Tree (DT) and Neural Network (NN) Algorithm
- combining GA (Genetic algorithm) and Decision Tree (DT) etc. ~[2-7]

The biggest challenge in most of these models however has been the inability to remove redundant/undesirable/wasteful features. Thus, this will the aspect that I'd focus in this project. Although I have proposed PCA/LDA as the primary aspects for feature selection, I will also spend more time understanding the effects of features from violin plots. Another idea that I am planning to explore is a computationally aggressive method, using ensemble methods guided by the results of feature selection. I am also considering the possibility of Gaussian Bayes over joint features, ie., by combining mean, se and largest values of each of the 10 features into one and building a Gaussian Naive Bayes model on top of that.

The kernel in this link https://www.kaggle.com/gargmanish/basic-machine-learning-with-cancer uses several classifiers such as decision trees, SVM, KNNs etc. and the best score they have come up with is 94% accuracy. This will be my benchmark, that is, I am going to try different approaches to beat this score.

1.6 Evaluation Metrics

Since the dataset is not very large, cross-validation should be applied, such that there is a enough data used for training as well as for testing. The various metrics that should be looked into are accuracy, precision, recall as well as F-beta scores, which is a single score metric combining the

precision and recall values. F1-score which is a harmonic mean of precision, recall might possibly a good metric in this case, to understand the balance between precision and recall obtained by the model.

Since the idea to test a variety of classifiers, it would be a good idea to compare the performances of each of the classifiers in terms of accuracy, F1-score as well times spent in training and testing.

Futher the hyperparameters of the classifiers can be optimized by maximizing the scores over a grid search. And there must be caution to avoid overfitting.

1.7 Project Design

Thus the overall workflow for this project would be:

1. Perform some preliminary data analysis
2. Explore the disributions of each of the features in each of the two classes after normalizing them
3. Explore the joint distributions of the features
4. Explore the possibility of combining radius, se, largest values of each of the 10 features into one and performing the above analyses based on that
5. Perform PCA, LDA to obtain a reduced feature space - for all features, and combined features from step 4
6. Explore ensemble methods that are a collection of weak learners, based on single features - guided by feature selection in the above steps
7. Use Gaussian Naive Bayes approach on the reduced feature space from step 3.
8. Test plane separation classifiers such as SVM, logistic regression
9. Test decision trees, random forests
10. Test neural networks
11. Compare the performances of the above classifiers, optimize the classifier hyperparameters and be wary of overfitting

1.8 References

1. Akay M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. 36, 3240-3247
2. Choi Y, Choi Y, Lee J, Nam J, Juhng S, Choi C. Analysis of Fine Needle Aspiration cytology of the breast: a review of 1,297 cases and correlation with histologic diagnoses. Acta Cytol. 2004;48(6):801–6
3. Maglogiannis I, Zafiropoulos E, Anagnostopoulos I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Journal of Appiedl Intelligence. 2007;30(1):24–36
4. Wang Y, Wan F. Breast cancer diagnosis via support vector machines.; Proc. 25th Chinese Control Conf; 2006. pp. 1853–6
5. Yang Z, Lu W, Yu R. Detecting false benign in breast cancer diagnosis. IEEE NNS-ENNS International, Joint Conf, Neural Networks. 2000;3:3655–8
6. Mu T, Nandi A. Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier. J Franklin Institute. 2007;344:285–311
7. Fuentes-Uriarte J, Garcia M, Castillo O. Comparative study of fuzzy methods in breast cancer diagnosis. Annual Meeting of the NAFIPS. 2008:1–5