Project Proposal- Breast Cancer

Predict if tumor is benign or malignant

Charles Ingulli & Yon Garber

2/24/2020

Introduction-

Breast cancer represents one of the leading causes of death every year. It is the most common type of all cancers and the main cause of death in women worldwide.

The breast cancer databases were obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. We found the dataset at UCI ML repository. The dataset contains information that was collected from a fine needle aspirate (FNA) procedure. When a patient is examined for breast cancer a FNA is obtained from the suspected area. Then the sample is observed under a microscope and assigned a value between 1-10 for 9 categories: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitosis. The dataset includes labels: 2 for benign and 4 for malignancy. In general, larger values indicate a greater likelihood of malignancy. However, no single category can determine if the sample is benign or malignancy.

If a patient's FNA indicates a benign sample then the patient is reexamined in 3 month and in a year or can choose to go through biopsy. A biopsy is 8 times more expensive than FNA and includes a surgical procedure but it tells for sure if the lump is cancerous or not.

Extensive literature exists on different methods that all aim to effectively classify a breast cancer tumor. Mangasarian et al. (1990) use a linear programming method to correctly classify tumors. A 30-dimensional vector containing tumor features such as nucleus area, radius, etc. is constructed to be used in classifying tumors between benign and malignant states through a separating plane. Vinod et al. proposed feature ensemble learning based on Sparse Autoencoders and Softmax Regression for classification of Breast Cancer into benign (non-cancerous) and malignant (cancerous). They used the same dataset as we are going to use. Their proposed method was assessed using various performance indices like true classification accuracy, specificity, sensitivity, recall, precision, f measure, and MCC. Their results were very promising- 98.6% true accuracy. In a little different aspect of breast cancer Chi et al. applied artificial neural networks (ANNs) to the survival analysis problem.

ANNs can easily consider variable interactions and create a non-linear prediction model and offer more flexible prediction of survival time than traditional methods. Another group of researchers tried to improve the classification accuracy by applying a new data selection method. Cai et al. proposed combining ensemble method and imbalanced learning technique for the classification of breast cancer data. First, Synthetic Minority Over-Sampling Technique (SMOTE), an imbalanced learning algorithm was applied to selected datasets and second, multiple baseline classifiers were tuned by Bayesian Optimization. Finally, they combined the optimized classifiers for the final decision. Another researcher, Akay, combined breast cancer diagnosis based on a support vector machine-based method with feature selection. He used the same dataset that we are going to use from the University of Wisconsin. He obtained classification accuracy of 99.51% for the SVM model that contained five features. In recent years there is a new trend which tries to fully automate the classification of images for breast cancer diagnosis. Kaymak et al. used two methods: Back Propagation Neural Network (BPPN) and radial basis neural networks (RBFN), to perform such diagnosis. Their accuracy was 59.0% and 70.4% respectively.

Initial Hypothesis-

Based on the existing literature for cancerous tumors, we hypothesize that malignant tumors will be classified as such based on their size and non-uniform shape. That is, malignant tumors will be large and abnormally shaped. Additionally, these same tumors are expected to be ones that divide rapidly and are constantly in a state of reproduction. Additionally, the data set contains variables on nuclei features. Malignant tumors should also have high levels of irregularity within these features.

Data-driven Hypotheses-

As the data is explored, we seek to understand which variables are significant in the classification of breast tumors. Additionally, it may be important to look at which variable values lead to a classification of malignant over benign. Based on an initial data exploration, large sizes of cells seem to contribute to a malignant classification. Nevertheless, it would be interesting to see the influence of each of the nine features on the diagnosis. Previous research showed that it is impossible to detect cancerous tumors from one feature but we can see the feature's influence on the prediction which might provide further medical insights. We don't know yet if the data is balanced or unbalanced but it would be interesting to apply different sampling techniques to reinforce our models.

Proposed work-

The goal of this project is to use the cell nuclei categories to classify a breast cancer tumor as benign or malignant. The methods of cross-validation will be used to create appropriate training sets that can be used to train the model as well as a test set to look at model accuracy. The methods of KNN, logistic regression, and quadratic and linear discriminant analysis to fit several models and comparisons will be made between the classification rates of each model. The use of cross-validation to take sampling errors out of the equation and to help us find the best model. One objective will be to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, and sensitivity. Using several different classification models and we can see which one performs the best. The usage of logistic regression will provide us further insight into the influence of the features relatively to the other features.

References-

1. O. L. Mangasarian and W. H. Wolberg: “Cancer diagnosis via linear programming”, SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
2. William H. Wolberg and O.L. Mangasarian: “Multisurface method of pattern separation for medical diagnosis applied to breast cytology”, Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: “Pattern recognition via linear programming: Theory and application to medical diagnosis”, in: “Large-scale numerical optimization”, Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
4. K. P. Bennett & O. L. Mangasarian: “Robust linear programming discrimination of two linearly inseparable sets”, Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).
5. Vinod, Jagannath Kadam, Manikrao Jadhav Shivajirao, and K. Vijayakumar. “Breast Cancer Diagnosis using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression.” Journal of medical systems 43.8 (2019): 1-11. ProQuest. Web. 16 Feb. 2020.
6. W. H. Wolberg, W. Nick Street and O. L. Mangasarian: “Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates.” 1994. [Online] 77 (2), 163-171.
7. Asri H, Mousannif H, Al Moatassime H, Noel T. “Using machine learning algorithms for breast cancer risk prediction and diagnosis.” Procedia Computer Science. 2016 Jan 1;83:1064-9.
8. A. Osareh and B. Shadgar, “Machine learning techniques to diagnose breast cancer,” 2010 5th International Symposium on Health Informatics and Bioinformatics, Antalya, 2010, pp. 114-120.
9. bin Othman M.F., Yau T.M.S. “Comparison of Different Classification Techniques Using

WEKA for Breast Cancer.” 2017. In: Ibrahim F., Osman N.A.A., Usman J., Kadri N.A. (eds) 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006. IFMBE Proceedings, vol 15. Springer, Berlin, Heidelberg

10. Chi CL, Street WN, Wolberg WH. Application of artificial neural network-based survival analysis on two breast cancer datasets. AMIA Annu Symp Proc. 2007;2007:130-134. Published 2007 Oct 11.

11. Tongan Cai, Hongliang He, Wenyu Zhang, Breast Cancer Diagnosis Using Imbalanced Learning and Ensemble Method, Applied and Computational Mathematics. Vol. 7, No. 3, 2018, pp. 146-154. doi: 10.11648/j.acm.20180703.20

12. Jos