CAPSTONE PROJECT PROPOSAL – *Machine Learning in Breast cancer diagnosis*

Christina Kallendorf

Motivation/Domain background

Machine Learning opens up a vast spectrum of medical applications. Some examples are medical image processing for diagnosis, classification of medical findings as well as disease prognosis and risk assessment [1]. Sidey-Gibbons claim that machine learning can be highly beneficial to the areas of diagnosis and outcome prediction in medical applications [2]. In particular, Machine Learning can be supportive to diagnose and classify specific sub-types of cancer from biopsy probes. In this context, Machine Learning can lead to an "effective and accurate decision making", as the authors of [3] point out. This may help to diagnose malignant tumors early and define the suitable treatment. As pointed out by [13], this may help to "increase survival rates from 56% to more than 86%". The project proposed here takes up one classification example of the area of cancer research and focuses on the identification of malignant breast tumors based on cytology features from biopsies.

Problem statement

Based on a medical patient's biopsy, it needs to be determined whether the investigated tissue contains malignant, i. e. cancer cells. In order to react to the growth of cancer cells quickly and appropriately, it is necessary to define such cells at an early state. Usually, biopsy results are classified by laboratory staff, based on their investigation of cell properties under a microscope. Therefore, a diagnosis is subject to individual judgement and human error [2].

Solution statement

A Machine Learning model can be able to identify breast cancer reliably for a given sample. The classification is itself based on a defined set of cytology features from biopsies, such as shape and size. As a result, cytology features obtained from biopsies enable a Machine Learning model to provide an effective, accurate and fast diagnosis. This has advantages over a subjective diagnosis generated by laboratory staff, see also [2], and may help physicians in decision-making, refer to [13].

Details on data set/input

For setting up a Machine Learning model, I employ results of breast tissue biopsies as test and training data. This data was collected in 1992 by University of Wisconsin Hospitals, Madison, by Dr. William H. Wolberg [4, 5, 6, 7] from tumor patients during the years of 1989 till 1991. It comprises eleven cytology characteristics of breast fine-needle aspirates. These features were extracted from collected images using methods developed in [4-6]. The data set is itself available at

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[8] and is available as a formatted .csv file at https://www.kaggle.com/johnyquest/wisconsin-breast-cancer-cytology-features. It comprises 699 data samples with the following 10 cytology features:

- ID number of the sample;
- Clump thickness,
- Uniformity of cell size,

- Uniformity of cell shape,
- Marginal adhesion,
- Single epithelial cell size,
- Bare nuclei,
- Bland chromatin,
- Normal nucleoli,
- Mitoses,

each of these ranging with values from 1 to 10, as well as

- Classification of benign cell, by value 0, and malignant, i. e. cancer, cells by value 1.

Evaluation metrics

Using machine learning for diagnostic classification, major objectives should be minimization of false positives and false negatives, while optimizing accuracy of the model. The authors of [2] suggest to optimize for the following metrics, aligned with medical standards:

- Sensitivity=true positives/actual positives = Precision
- Specificity=true negatives/actual negatives
- Accuracy= (true positives + true negatives) / total predictions.

Research on this subject matter/ Bench mark model

This data set has often been employed for machine learning classification studies. For example, refer to references [8-13]. As a benchmark, I have selected an SVG model that was presented by [2]. The authors were able to achieve the following evaluation metrics using an SVG model:

- Sensitivity = 0.97
- Specificity = 0.94
- Accuracy = 0.96

I would like to compare the performance of an XGBoost algorithm with the given model.

Project Design

I would like to conduct this project by the following steps:

1. Explore and prepare data (drop duplicates and null values). Investigate on data by visualization (scatter plot of each two features, bar diagram for each feature).
2. Check for correlations of the single feature values, as a cell constitutes an organism with interdependent components. If features have high correlations, these features will be dropped.
3. Perform a PCA analysis on the data in order to further reduce dimensionality (i. e. features).

4. Shuffle data and partition it into test and training data sets. Separate label (i. e. the classes of the sample). Upload this data in .csv format to S3 bucket.
5. Build, deploy and optimize the model:
    a. Define a binary XGBoost classifier and train it on the training data set,
    b. Deploy the model,
    c. Predict class of test data and compute the evaluation metrics suggested above..

    Repeat these steps and tune the hyperparameters of the model for optimization.
6. Clean up resources.

Bibliography

[1] de Bruijne, Marleen (2016). Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33. doi: 10.1016/j.media.2016.06.032

[2] Sidey-Gibbons, J., Sidey-Gibbons, C. (2019). Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*, 19(64). doi: 10.1186/s12874-019-0681-4

[3] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, Volume 13, 2015, pp. 8-17. doi: 10.1016/j.csbj.2014.11.005.

[4] William H. Wolberg and O.L. Mangasarian (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, U.S.A*., 87, 9193-9196.

[5] O. L. Mangasarian and W. H. Wolberg (1990). Cancer diagnosis via linear programming. *SIAM News*, 23 (5), 1-18.

[6] 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.

[7] K. P. Bennett & O. L. Mangasarian (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software,* 1, 23-34.

[8] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004.

[11] Hussein A. Abbass. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Intelligence in Medicine, 25. 2002.

[12] Agarap, Abien Fred M. (2019). On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. arXiv:1711.07831v4 [cs.LG] 7 Feb 2019

[13]Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. (2016). Machine learning models in breast cancer survival prediction. *Technol Health Care,* 24(1), 31-42. doi:10.3233/THC-151071