Breast Cancer Diagnosis with Machine Learning

Zhe Zhou

Beijing Royal School, Beijing 102200, China

Abstract. An early-stage breast cancer diagnosis usually results in a high survival rate within five years. This research seeks the feasibility of applying a specific machine learning algorithm, logistic regression, to a public dataset. The result is two models, reaching accuracies of 0.965 and 0.936. It confirms that the generality of applying machine learning algorithms is beneficial to medical diagnosis. Even though this is not the first research to exploit the application of machine learning in the medical field, it still provides some valuable insights for future research.

Keywords: Breast cancer classification; Logistic Regression; Machine Learning.

1. Introduction

Machine learning utilizes a large amount of data and repetitive mathematical calculations to learn the trend and weights of different parameters of a dataset. Due to the explosive growth of data on the Internet and the improvement of hardware performance in recent years, machine learning can thus be realized and applied in various common-use fields. [1] One of the most meaningful ways of using it is in the medical area, where a manual diagnosis cannot perfectly handle unconstructed data such as images and abstract numerical values. However, that is what machine learning can do, excavating the potential value of those unconstructed data. [2] Recent advancements in machine learning in medical areas such as heart disease [3] indicate that applying machine learning to diagnose other conditions can be feasible. Compared to existing research using machine learning to classify benign and malignant tumors [4][5], this research aims to achieve high accuracy with less restriction on input data, building a more straightforward and practical model.

Cancer is a disease caused by the uncontrolled replication of cells in the human body. The risk of developing cancer increases exponentially with age; people's life span has been extended tremendously since the last century due to the continuous development in the medical field. So, many more people suffer from the illness of cancer as a result of increasing age. More specifically, malignant tumors indicate cells that spread to other body parts, while benign tumors show cells that stay in their place. Telling whether a tumor is malignant or benign can determine cancer and the following treatments since benign tumors are often less harmful. [7]

One type of cancer for women is breast cancer. It has been recorded as the deadliest cancer for women worldwide, producing an estimated 1.67 million new cases in 2012 and causing 522,000 deaths. [6] Taking a specific view of the statistics, it is clear that the 5-year survival rate for women in developing countries (e.g., China) is 73%, compared to nearly 90% in developed countries (e.g., the United States), mainly because of early screening for breast cancer is more prevalent in developed countries. [8] [9] A reasonably precise model requiring a small amount of diagnostic data takes advantage of artificial intelligence with lower cost and higher accuracy, making an early diagnosis as prevalent as possible in more countries. As a result, the research consists of two models. One model will build upon all the available features given in the dataset. Then, revising the previous model will create another model with fewer features essential to prediction accuracy.

2. Materials and Methods

Description of data. UCI Machine Learning Repository provided data in 1995. It consisted of 30 numerical features computing from the digitized image of a fine needle aspirate (FNA) of a breast mass of 569 real-life instances. Those features were selected based on Multisurface Method-Tree [10]. Within the dataset, 212 cases were diagnosed as malignant, and the rest 357 cases were diagnosed as benign.

Table1 Sample data

id	diagnosis	radius_mean	texture_mean	…	symmetry_worst	fractal_dimension_worst
842302	M	17.99	10.38	…	0.4601	0.1189
842517	M	20.57	17.77	…	0.275	0.08902
84300903	M	19.69	21.25	…	0.3613	0.08758
84348301	M	11.42	20.38	…	0.6638	0.173
84358402	M	20.29	14.34	…	0.2364	0.07678

There were only ten main parameters, and the mean, the standard error, and the worst (mean of the three largest values) of each parameter were displayed separately.

Logistic Regression. Logistic Regression was widely used for binary classification and prediction based on a given set of independent features. It outputted the possibility and classified each instance based on a threshold. After each iteration of calculating the performance of the current weights of each parameter, the model will adjust those weights based on how they affect the outputting probability. As more and more iterations are finished, the weights would tend to fit into an optimal coefficient for our dataset by gradient descent. Then the performance in the real-life environment was evaluated by calculating the accuracy of predicting the test dataset. [11] In this research, I used the scikit-learn implementation.

As said in the introduction, the first model was built upon all 30 valid features. Splitting training and test sets with a ratio of 7 to 3, the training data was then normalized to make it easier for our model to converge. With a maximum of 5000 iterations and an l2 penalty, the resulting model reached an accuracy of 0.965. The second model was created under the same training condition but this time, some features were removed if they are closely related to other features. In another word, they were not independent enough to provide the model with useful information to classify malignant and benign tumors. A heatmap was created (Supplementary S1). Features with an absolute value of correlation greater than 0.70 were removed, and 10 features remained. The resulting accuracy reached 0.936.

3. Discussion

In this study, I show one of the machine learning approaches, logistic regression, to classify the malignancy of a tumor based on its numerical features. The result indicates the high accuracy of both models (0.965 and 0.936) in predicting the test set. Even a highly accurate method, Digital Mammography, in detecting breast cancer would result in only a 0.893 accuracy. [12] The difficulty of applying a machine learning model usually comes from the dataset since the result of the model is extremely specific to the dataset I used. Future validation with another different set of clinical data under a restricted condition can further expand the generality of the resulting model.

Another noticeable problem is that applying a logistic regression algorithm under a small amount of data quickly fails to converge due to highly complex features. So, sometimes, reducing the number of features used based on the correlation map can help simplify the model's complexity, making it easier for us to build a feasible model with high accuracy. It also indicates a lower cost of the diagnosis, so people might afford to attend a simplified test to see if they have breast cancer, and this is also the primary goal of this research.

Besides, this research confirms a possible general way of applying machine learning algorithms in medical diagnostics. With a large amount of data, other algorithms like random forest will also work better on this classification problem if the data is well constructed. A neural network can be another competitive choice, but only applying it to image data can be more advantageous.

References

[1] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

[2] Attaran, M., & Deb, P. (2018). Machine learning: the new “big thing” for competitive advantage. International Journal of Knowledge Engineering and Data Mining, 5(4), 277. https://doi.org/10.1504/ijkedm.2018.095523

[3] Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 128, 102289. https://doi.org/10.1016/j.artmed.2022.102289

[4] Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. 2010 5th International Symposium on Health Informatics and Bioinformatics. https://doi.org/10.1109/hibit.2010.5478895

[5] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT). https://doi.org/10.1109/ebbt.2018.8391453

[6] Stimpfel, M., & Virant-Klun, I. (2016). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Journal of Cancer Stem Cell Research, 4(3), 1. https://doi.org/10.14343/jcscr.2016.4e1003

[7] Sinha, T. (2018). Tumors: Benign and Malignant. Cancer Therapy & Oncology International Journal, 10(3). https://doi.org/10.19080/ctoij.2018.10.555790

[8] Zeng, H., Zheng, R., Guo, Y., Zhang, S., Zou, X., Wang, N., Zhang, L., Tang, J., Chen, J., Wei, K., Huang, S., Wang, J., Yu, L., Zhao, D., Song, G., Chen, J., Shen, Y., Yang, X., Gu, X., . . . Yu, X. Q. (2014, October 3). Cancer survival in China, 2003-2005: A population-based study. International Journal of Cancer, 136(8), 1921–1930. https://doi.org/10.1002/ijc.29227

[9] Cancer Facts & Figures 2010 | American Cancer Society. (n.d.). Retrieved September 7, 2022, from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2010.html

[10] Bennett, Kristin P. (1992). Decision Tree Construction Via Linear Programming (No. TR1067). The University of Wisconsin-Madison. https://minds.wisconsin.edu/handle/1793/59564

[11] Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019, October). Logistic Regression Model Optimization and Case Analysis. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). https://doi.org/10.1109/iccsnt47585.2019.8962457

[12] Zeeshan, M., Salam, B., Khalid, Q. S. B., Alam, S., & Sayani, R. (2018, April 8). Diagnostic Accuracy of Digital Mammography in the Detection of Breast Cancer. Cureus. https://doi.org/10.7759/cureus.2448