Alex Trexler and Alex Trunzo
CISC484

Comparison of Machine Learning Techniques for Breast Cancer Diagnosis

Abstract

Breast cancer is the most common form of cancer in the world. One of the most important factors in treating breast cancer is early detection. As such, it is imperative that we have multiple approaches available to quickly and accurately diagnose it. This paper examines Machine Learning models as an important tool that doctors can utilize in aiding with the diagnosis of breast cancer. Two machine learning techniques, decision trees and neural networks were trained and tested on a breast cancer dataset. The results were compared to one another in an attempt to determine which model was better suited for the task. This information could be used to enhance our understanding of the possible approaches for detecting breast cancer, thereby helping with the diagnosis of people who may be suffering from it. By comparing common decision tree and neural network algorithms, it was determined that the neural network approach has an advantage in overall accuracy. This can be complicated by variance in training and testing, but our neural network classified the data more accurately on average. There were also differences in the levels of recall and precision between the two models, which are important factors to consider in the context of a medical diagnosis.

Introduction

Breast cancer is the most prevalent form of cancer worldwide. There are various different types of breast cancer, which are defined by where exactly the cancer starts within the breast. Some women are at a higher risk of developing breast cancer than others. One of the main risk factors for breast cancer is dense breasts. Women with dense breasts are more likely to get breast cancer than others. Some other risk factors include age, genetic mutations, and family history. [1]

Like other cancers, breast cancer can metastasize, or spread to other parts of the body. Breast cancer has four different stages describing how far it has spread. Stage 1 describes cancer that has spread to tissue and inhibits a small area, stage 2 describes cancer that has spread to a larger area and may involve lymph nodes, stage 3 describes cancer that has spread to a large region involving more lymph nodes, and stage 4 describes cancer that has spread to other parts of the body. [2] The five-year survival rate for stage 1 and stage 2 breast cancer is 99.3%, for stage 3 breast cancer it is 86.3% and for stage 4 breast cancer it is 31%. [3] As the stages get higher,

treatment of breast cancer becomes much more difficult and extensive, and as a result, the success rate declines immensely. As a result, it is extremely important that breast cancer is detected as early as possible, in order to ensure that the patient has a positive outlook and the best chance of recovering.

By examining the effectiveness of various machine learning methods in diagnosing breast cancer, we aim to provide doctors with another tool they could use to detect breast cancer early and increase the odds of successful treatment. In Section 2, we describe the machine learning models we used, which were decision trees and neural networks. We describe some related works and how our approach relates to and differs from them in Section 3. In Section 4 we described our methods, which involved training and testing a decision tree and neural network on the original dataset and then running 200 testing trials with each of the models. In section 5 we present and compare the results to each other to determine which model was more accurate.

Related work

Machine learning is becoming a more prevalent tool in the medical field year after year. Other recent papers have explored various ML models for medical diagnosis. Naji et al. [6] looked at 5 different simple models to predict breast cancer but did not test neural networks. Mahoto et al. [5] explored different machine learning models as well as bagging and ensemble methods. They found that Multilayer Perceptron models work well on heart disease and eye infection data. Das et al. [4] designed a specially trained system to classify breast cancer data, with the goal of addressing the class imbalance problem. They found a K-Means approach to best address the issues. It is clear that research in this area is evolving quickly and is likely to save lives as long as models can be improved upon. Our goal with this project was to determine how much of an effect different models can have on the accuracy of medical diagnosis.

Methods

The two machine learning models chosen for comparison were the decision tree and the deep neural network. These models were chosen based on their popularity as well as differences in complexity and processing power. Decision trees are simpler models that use a tree-shaped graph to classify inputs. The goal of a decision tree is to create a predictive model that can make predictions by following a path from the root node to a terminal node of 0 or 1. At each “branch” in the tree, the data is split based on a single characteristic. The splits can be computed with various decision formulas such as entropy or information gain, which measure the uncertainty and reduction of entropy respectively. At the end of these branches, the data is classified based on the path it has taken. In this way, the classifier can identify medical diagnoses by cross-referencing data parameters. This structure provides clear insights into the decision-making process, as each path from the root to a leaf represents a set of conditions leading to a particular outcome, so it is easy to see what is contributing to the decision.

The other model was a deep-learning neural network model, which is based on a simulated brain-like network of activation nodes. The data points are fed into input nodes and then that data is processed through a series of interconnected nodes that each affect what is passed on to the next layer. Activation functions, such as ReLu, contribute to these decisions made in the hidden layer. The result is a prediction of the chance of each outcome, of which the most likely is chosen. The neural network is a more complex model than the decision tree and is not as easily understood after training has been done. Unlike a decision tree, it is not exactly clear what went into each prediction.

Experimentation

The dataset used for this project is a possible breast cancer measurement and diagnosis dataset. The dataset contains information and measurements about "suspicious lumps", including mean radius, perimeter, texture, area, smoothness, and whether those lumps were found to be cancerous or benign. The dataset contains 569 data, 357 of which are positive (malignant) instances, and 212 of which are negative (non-cancerous) instances. Each data instance has a value for all of the 6 attributes, there is no partially completed data. In addition to this, there is no duplicated data. A 70/30 split ratio was used for training and testing. 399 data points were used for training, and 170 data points were used for testing. Random sampling was used to decide which category the data was split into. The table below shows an example of some data instances.

Table A. Example of Data Instances from Breast Cancer Dataset

mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
12.45	15.7	82.57	477.1	0.1278	0
13.54	14.36	87.46	566.3	0.09779	1

The "scikit-learn" machine learning package in Python 3 was used for training and testing because of its ease of use. The library contained a default model for both methods that was modified slightly for the task.

The data was split into training and testing sets of 70:30, using random sampling. Then the data was normalized, and feature scaling was performed to ensure that differing scales wouldn't impact the performance of the neural network. Then the decision tree was trained using the fit method on DecisionTreeClasifier from "scikit-learn" with "entropy" selected as the criterion parameter (all other parameters were default). This parameter sets the quality of a split to be measured using the entropy metric as previously described, this is also called Shannon Information Gain and is referred to as such in the documentation for the parameter. After this, the

neural network was trained using the fit method on MLPClassifier from “scikit-learn”. Default parameters were used with hidden_layer_sizes set to (10, 10, 10) and max_iter set to 500. Both of the models were then tested using the predict method with the testing set. Lastly, a confusion matrix was calculated, keeping track of the true positives, false positives, true negatives, and false negatives for each of the models. The entire process previously described was repeated 200 times, to get a good understanding of how the models will perform against each other on average, as well as to minimize the impact of chance variance on the results. The average of true positives, true negatives, false positives, and false negatives for the 200 trials was calculated, then the average accuracy for each model was computed by adding the average percentage of true positives plus the average percentage of true negatives. The precision was calculated by taking the total number of true positives and dividing it by the total number of positives. The recall was calculated by taking the total number of true positives and dividing it by the total number of true positives plus false negatives. The section below shows the results.

Results

The figures below show the average results for the 200 trials.

Figure A. Decision Tree Results

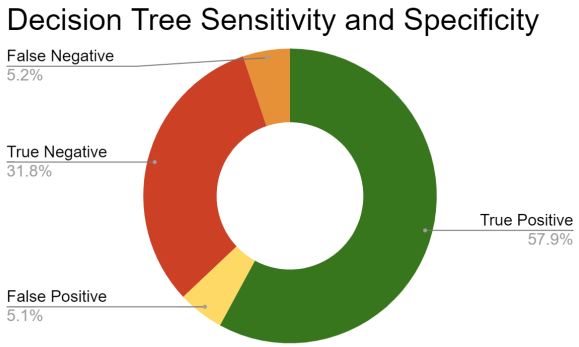

Figure B. Neural Network Results

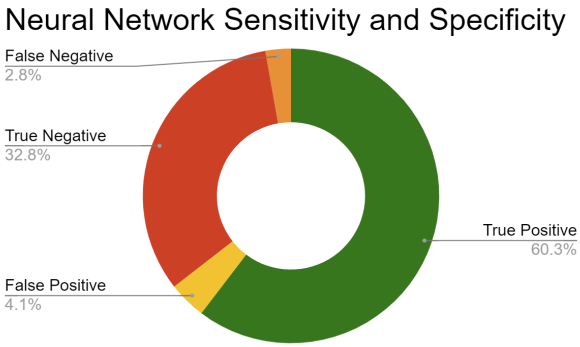

Table B. Accuracy, Recall and Precision of the Models

	Accuracy	Recall	Precision
Decision Tree	89.7%	92.0%	91.8%
Neural Network	93.1%	93.6%	95.6%

The accuracy of the neural network beat out the accuracy of the decision tree model by a little over 3 percentage points. Additionally, the neural network achieved a much better recall than the decision tree. Recall is a measure that compares true positive cases with the total number of positive cases (true positives plus false negatives). In the context of a breast cancer diagnosis, false negatives should be eliminated whenever possible because they could lead to a patient falsely believing they are cancer-free when they do in fact require treatment. This could significantly delay the detection of cancer and subsequently the treatment. The neural network's average recall was 95.6%, which is significantly better than the decision tree score of 91.8%.

False positives are also a concern as they can contribute to unnecessary stress. Precision is a measure of the rate of true positives to total positive predictions (true positives plus false positives). The neural network had a lead in precision as well, but lesser so, at 93.6% to 92%. Positive diagnoses will usually come with second opinions, so the consequences are not as disastrous as a false negative. Therefore it is beneficial that the neural network appears to concentrate its improvements on increasing recall. The decision tree model doesn't perform badly, however, the data suggests that the neural network approach is better suited for this particular problem.

Conclusion

This project investigated the effectiveness of two different machine learning models on breast cancer detection. Normal decision tree and neural network models were compared using the same dataset. 200 trials were run to account for variance in data selection and model initialization. The confusion matrix of each model was compiled and analyzed for trends. The neural network model performed marginally better than the decision tree model. It had an average accuracy improvement of 3.4% over the decision tree model, producing an average of 2.4% fewer false negatives and 1.1% fewer false positives. This evidence shows that model selection is crucial to achieving maximum accuracy in medical diagnosis. Only two model types were tested, so it is likely that other implementations may produce different results. Further comparisons are needed to test the neural network approach against other models. Variations of the neural network model should also be tested, such as altering the number of hidden layers.

Neither model had a high enough accuracy to be fully trusted on its own, however, it can be used in the process of a professional medical diagnosis. Machine learning models are already being used in various areas of medical diagnosis for their speed and accuracy. Medical professionals can consult models such as the ones explored as a second opinion. If medical professions and machine learning models agree on a diagnosis, it is more likely to be accurate. If the models and medical professionals differ in their judgment, it can be a sign to reexamine evidence or perform more testing to avoid a misdiagnosis.

References

[1] CDC (n.d.). What Is Breast Cancer? Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

[2] MSKCC (n.d.). *Stages of Breast Cancer*. Memorial Sloan Kettering Cancer Center. https://www.mskcc.org/cancer-care/types/breast/diagnosis/stages-breast?utm_campaign=It+Takes+MSK+GM&utm_medium=Paid+Search&utm_source=Bing&utm_content=345583486&utm_term=345583486&pn_mapping=pn_13#what-is-stage-0-breast-cancer-

[3] Welsh, J. (2023, October 22). Stages of Breast Cancer: Understanding the Classification System. Verywell Health. https://www.verywellhealth.com/stages-of-breast-cancer-8362458

[4] Das, A. K., Biswas, S. K., Mandal, A., Bhattacarya, A., & Sanyal, S. (2023). Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP). Expert Systems With Applications, 242. https://doi.org/10.1016/j.eswa.2023.122673

[5] Mahoto, N. A., Shaikh, A., Sulaiman, A., Reshan, M. S. A., Rajab, A., & Rajab, K. (2023). A machine learning based data modeling for medical diagnosis. Biomedical Signal Processing and Control, 81. https://doi.org/10.1016/j.bspc.2022.104481

[6] Naji, M. M., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science, 191, 487-492. https://doi.org/10.1016/j.procs.2021.07.062