Predicting Breast Cancer Survivability Using Deep Learning and Support Vector Machines

Marian Qian
TJHSST
May 2021

Overview

This project evaluates whether combining a feedforward neural network and a radial basis function or linear support vector machine will help the prediction of breast cancer survivability. Breast cancer survivability is how long the patient survives after their initial diagnosis, and understanding survivability is crucial to aiding healthcare physicians in providing both effective and efficient treatment plans to patients. Previous papers have addressed this problem by using machine learning methods such as neural networks and support vector machines as individual algorithms, but this project investigates whether combining these two methods will improve the accuracy of survivability prediction. While the results of this project did not show benefits or an increased accuracy from combining a support vector machine with a neural network, a smaller standard deviation of model performances was shown in the radial basis support vector machine combined with neural network.

Motivation

Breast cancer is one of the most common cancers, not only in the United States but across the globe. Over 250,000 women in the US are affected by breast cancer each year, and by 2025, the expected number of cases will rise to approximately 19.3 million. The prognosis of breast cancer, which describes the behavior of the disease after the initial diagnosis, mainly consists of recurrence, whether cancer will come back, and survivability, how long the patient will remain alive after the initial diagnosis. Understanding these factors of prognosis would help guide treatment plans, such as whether a patient needs chemotherapy or hormone therapy, and proper follow-up methods that will most likely benefit the patient's survival rate of cancer. Knowing patient survivability would allow patients to avoid unnecessary treatments and surgeries, such as biopsies, and reduce hospital costs, so machine learning methods have been used to predict these survivability outcomes that result in high accuracies.

Background

Previous studies have investigated the use of machine learning methods, including support vector machines, decision trees, and random forests, to predict breast cancer survivability. The way that we can define survivability is whether the patient is still alive after a specified period of time -- this means that they survived and are still "alive". These studies have chosen 5 years or 60 months as the survivability time frame, which means that if the patient lived longer than 5 years, then they count as having survived the cancer. This five-year threshold will be used to determine which patients are in which classes.

The reason why an SVM is used is because several studies have experimented using SVMs as the last layer of a convolutional neural network (CNN), as they often perform with higher accuracy and are better classifiers, which is why this project combines an SVM with a regular feed-forward network.

Methodology

Data was used from the Surveillance, Epidemiology, and End Results (SEER) Cancer Dataset from the National Institute of Health, where there were a catalog of different cases of cancer tumors from the year 2000 to 2017. Cases with missing values were removed, and the continuous variables were normalized while the categorical variables were encoded. The majority class was randomly resampled in order to have the same number of cases for each class of whether the patient was alive or dead.

Three different model structures were trained each with the same training configurations.

Regular Neural Network: The regular feed-forward neural network consisted of 3 hidden dense layers with 200, 150, and 100 nodes respectively. The last layer consisted of one node that gave the prediction for which of the two classes the patient would be in (whether they were alive or dead). The rectified linear activation function, or ReLU, was used as the activation function after each dense layer, which changes all of the negative output values from the neural network to 0. Dropout layers with a 30% dropout rate and batch normalization layers were inserted after dense layers as well to help with regularization and to prevent overfitting. In addition, binary cross entropy loss was used, as it's the most common loss function used for binary classification problems.

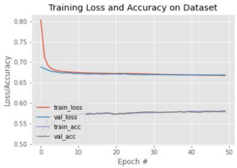
Linear SVM with Neural Network: In order to implement the linear SVM in the last layer, the loss function was changed to hinge loss during training and a kernel L2 regularizer was added to the last layer with one node.

RBF SVM and Neural Network: The RBF SVM in the last layer of the neural network was implemented by adding a layer that projects the inputs into a higher dimensional space. As complex data is plotted onto a higher dimensional space, their separator to classify the data becomes less complex or more "linear" in structure which then makes it easier to differentiate the classes. The number of dimensions for the layer's output was set to 20, and the parameters of the layer were randomly resampled from the Gaussian distribution.

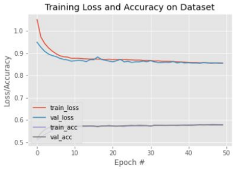
Training: All models were trained with 1cycle learning rate with a maximum value of 0.001, Adam optimizer, and through 50 epochs with 128 batch size. The weights were initialized using He initialization.

Results

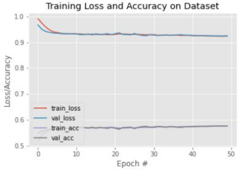
Using a neural network with these parameters and 3-fold cross validation, the model was able to reach an average accuracy of 57.85%, 57.73% when combined with a linear SVM, and 57.39% when combined with the RBF SVM. The graphs below plot the training and validation accuracy and loss of the best run for the three models.



Graph 1. Regular Neural Network.



Graph 2. Linear SVM with Neural Network.



Graph 3. RBF SVM with Neural Network.

Conclusions

Due to time constraints, the accuracy of the neural network was not as high as originally intended; however, the highest accuracy for previous studies that used the same SEER dataset and a neural network reached an accuracy of 65% (Park et al., 2013, Kim et al., 2013). Although the model wasn't able to reach a higher accuracy, it was still trained to a standard relatively close to that of previous projects.

Looking at the results of the neural network and SVM combined models, they did not improve the accuracy of the neural network as originally hypothesized. This could be due to how the data used was not complex enough, as some data was missing from the SEER dataset that was used in other papers, such as marital status and whether the patient was exposed to radiation (Park et al., 2013). In addition, the lower accuracy of the normal neural network could have influenced or worsened the performance of the model when it was combined with a linear and RBF SVM.

Another possibility for the decreased accuracies could simply be that the SVM is redundant or not needed with the neural network specifically for this problem. Previous papers used CNNs with SVMs, and this project used a normal neural network; the non linear transformation the SVM gives is already accounted for in the neural network activation functions, such as ReLU, so this additional layer could have messed with the output of the network when it wasn't needed.

Despite the lower accuracies, neural network and SVM combined structure can still be useful on this problem given more work on improving the accuracy of the normal neural network and trying more hyperparameter searches regarding the number of layers, number of nodes, and the learning rate of the normal neural network.

Additionally, one result from this project was that the combined linear SVM model had a smaller standard deviation, which could indicate that the models are more consistent, though the difference in standard deviation between the models is very small.