USING MACHINE LEARNING TO PREDICT PATHOLOGICAL COMPLETE RESPONSE AND RELAPSE-FREE SURVIVAL OUTCOMES TO IMPROVE PATIENT STRATIFICATION AND TREATMENT

Jeffrey Adjei, Euan Deas, William Sephton, Simon Stubbs, Grace Otuagomah

Univeristy of Nottingham
Nottingham, United Kingdom

ABSTRACT

Index Terms—

1. INTRODUCTION

1.1. Background

Breast cancer is the most prevalent cancer among women in the UK. Chemotherapy is a common treatment strategy used to reduce the size of locally advanced tumours before surgery. However, this approach is not always effective, as only 25% of patients achieve a pathological complete response (PCR) at surgery, which is associated with a higher likelihood of cure and longer relapse-free survival (RFS). The remaining 75% of patients experience residual disease, leading to varied prognoses. Early prediction of PCR and RFS before chemotherapy could enable better patient stratification and more personalized treatment plans.

1.2. Aim

The aim of this project is to develop machine learning models to predict PCR (a classification problem) and RFS (a regression problem) using pre-chemotherapy clinical and MRI-derived features. These models will leverage a dataset derived from the I-SPY 2 Trial, which includes 11 clinical features, and 107 MRI-based features extracted from tumour regions.

1.3. Objectives

1. Data Preprocessing: Handle missing values, outliers, and normalization to prepare the data for model development.
2. Feature Selection: Identify the most predictive features, ensuring key features (ER, HER2, and Gene) are retained in the models.
3. Model Development & Hyperparameter Tuning:: Build and evaluate machine learning models for PCR classification and RFS regression using appropriate algorithms. Optimize model performance by tuning hyperparameters using grid search or other techniques.
4. Model Testing and Evaluation: Evaluate model performance on a reserved test set using balanced classification accuracy for PCR and mean absolute error for RFS.
5. Code and Submission: Prepare final code for testing on a hidden dataset and submit prediction results.

This project can then be used to develop predictive tools that can aid clinicians in making more informed treatment decisions for breast cancer patients.

2. RELATED WORK

3. METHOD

The ML Pipeline Overview; The general machine learning pipeline is well known. If the data is unclean it must be cleaned, which often involves replacing missing values with appropriate means or medians. The data can be subject to optional standardisation/normalisation and to optional dimension reduction. Different types of ML models are considered. Hyperparameters are chosen and tuned. Provisional assessments of models, hyperparameters etc. are conducted through strategic use of the available data. The process should end with the selected model or models with the selected hyperparameter being trained on all the available data.

3.1. Data Cleaning

Missing values in the data where all replaced with the features' medians. (This meant that for binary features - such as 'pCR (outcome)' - the missing values will have been replaced by 1 or 0 depending upon which had the higher frequency.)

3.2. Normalisation

Each data point was rescaled so that when all the features, excluding the key features of 'ER', 'HER2', and 'Gene' and the target features of 'pCR Outcome' and 'RelapseFreeSurvival (outcome)', are considered as a vector have an 'l2' norm of one.

3.3. Dimension Reduction

All features excluding 'ER', 'HER2', 'Gene', 'pCR Outcome' and 'RelapseFreeSurvival (outcome)' were subjected to Principal Components Analysis, whereby the number of components were selected so that the amount of variance to be explained was greater than 95%.

3.4. Hyperparameter Tuning

The hyperparameters of both the SVR and SVC were selected from a pre-set parameter grid through nested cross validation.

3.5. Training

SVR and SVC models with those selected hyperparameters were then trained on the entire data set.

4. EVALUATION

The dataset used to train the model was a simplified version of the public dataset from The American College of Radiology Imaging Network I-SPY 2 TRIAL **??**. This data set was generated to evaluate the effectiveness of quantitative diffusion weight imaging (DWI) for assessing breast cancer response to neoadjuvant chemotherapy (NAC). Of the 400 patients in the dataset each had 118 features. 11 of those features are clinical, describing the patients age, tumour stage, lymph node status and more. 107 of those features are MRI-based, extracted from the tumour region using radiomics feature extraction package. Apart from the age all clinical data was categorical, and all MRI-based data was continuous. The dataset also includes two outcomes, a binary pathological complete response (PCR) classification and a continuous relapse-free survival (RFS) time. The dataset comes with a number of different challenges that need to be tackled before training a model on it. One of these challenges is that during feature selection it is important that the ER, HER2 and Gene are retained as they are very important features. Another challenge with this dataset is that the PCR outcome and almost all the clinical features contain missing data. Particularly the Gene feature was missing in the cases of 88 patients. This was dealt with by trying a number of data imputation methods. The biggest challenge of this dataset is the low number of patient samples, this paired with the large number of features means that training and validating the model poses a significant risk of over-fitting, leading to poor generalisation on unseen data. This sample number is also unbalanced with more samples for when PCR is not possible. To deal with this cross validation was used, but due to the sample number this also poses challenges as the number of folds must be balanced. From our testing the support vector classification model performed best for PCR classification. The cross fold validation showed a mean accuracy of 62.25%, with the best fold performing with a 79% accuracy and the worst performing with 23%. From the cross fold validation it showed that the best performing parameters were balanced class weight, the Radial Basis Function kernel, a C of 1 and a gamma of 0.01. By retraining the model with these parameters on an 80:20 test split we obtained an accuracy of 75%. The model is much better at predicting when a PCR is not possible (83% f1 score), with a f1 score of 50% for when it is possible. This is also shown in the confusion matrix in Figure XX. The confusion matrix also shows the class imbalance in the dataset. Figure XX shows the ROC curve for the model that indicates that the model has acceptable performance in distinguishing between the classes being moderately better than random guessing, however it is not exceptionally strong. Based on the tested regression models, we chose the support vector regressor for the RFS regression. The cross fold validation showed an average mean squared error (MSE) of 739.18, an R-squared score of -0.0074 and a mean absolute error (MAE) of 21.2267. This indicates fairly poor performance. From the cross fold validation the best parameters were the Radial Basis Function kernel, a C of 1 and an epsilon of 0.2. Training the model with these parameters on an 80:20 split of the dataset gave an MSE of 797.89, R-squared of 0.0002 and MAE of 21.64. Following on from the cross fold validation, the model continues to struggle in predicting values, with a lack of precision and performing almost identically to simply predicting the mean of the target variable for all data.

5. DISCUSSION

Both models used have their own advantages and disadvantages that are crucial in understanding the choice to apply them in this scenario. The Support Vector Classification (SVC) model is extremely effective when dealing with high dimensional spaces which in turn allows for a solid ability to generate complex decision boundaries, which is important with a dataset of this style. With the added ability to use margin maximization, SVC models generalize efficiently when using unseen data increasing its application potential in terms of versatility. The SVC model has some disadvantages such as an enhanced focus required on defining the hyperparameters for model accuracy, if the incorrect kernel is selected then the output will be heavily impacted. SVC models also have the downside of being computationally expensive, meaning on unknown datasets of varying sizes, results may take longer to process contrasting to other machine learning methods. The advantages of SVC models outweigh the disadvantages as effort has been dedicated into tuning the hyperparameters and being provided with an example dataset allows the feature selection process to be smooth and accurate. The Support Vector Regression (SVR) model has a specific feature that makes it a powerful option, the robust ability to ignore errors within a margin of tolerance to focus on general trends rather than being affected by every outlier. Along with SVC models, SVR models also work effectively with high dimensionality increasing the number of applications and scenarios the model is applicable in. The SVR model does not just share positive similarities with SVC models, but also the negatives that come with being Support Vector Methods. One of these negatives is the computational complexity, especially with large datasets, meaning depending on the dataset results could take longer to be produced which could be an issue if the scenario requires efficiency over accuracy. The need for essential hyperparameter tuning also plagues SVR models but fortunately, as stated previously, time has been dedicated into researching and analysing the best hyperparameter settings to ensure these models are the correct and appropriate models for the dataset provided.

6. CONCLUSION

7. REFERENCES