COMP3009 Machine Learning 2022-23
Assignment 2

Jarrad Foley, Stephen McSweeney, Ahmadreza Omidvar, Ameya Pimpley, Aryavrat Singh

ABSTRACT

Breast cancer is the most common type of cancer in British women. Chemotherapy is a commonly used therapeutic strategy to reduce the size of locally advanced tumors before surgery. However, chemotherapy is a toxic process for the human body and is not always effective. Complete tumor resolution with surgery, known as Pathological Complete Response (PCR), is likely to achieve a cure and longer Recurrence-Free Survival (RFS). RFS is the length of time a patient's life without signs or symptoms of their cancer after stopping treatment for their primary cancer. However, only 25% of patients receiving chemotherapy achieve PCR, and the remaining 75% have the residual disease and variable prognosis. Pre-chemotherapy information could be used to predict PCR and RFS to improve patient stratification and treatment. The main aim of this study is to use advanced machine learning techniques, both clinically measured characteristics and characteristics obtained from Magnetic Resonance Imaging (MRI) to identify PCR and RFS.

Using SVM, Logistic Regression, Decision Trees, and Neural networks getting a best results of 84% for classification of PCR and getting a best mean absolute error of 21.27 for regression of RFS.

1. INTRODUCTION

Cancer is one of the main causes of loss of life worldwide, and breast cancer being most commonly diagnosed in women, "with more than 2 million women being diagnosed and 685,000 dying each year" [1]. "Breast cancer outcomes are related to the stage of diagnosis, and early diagnosis is associated with improved survival" [2].

The underlying principle of breast cancer screening is that early detection of the disease and establishment of effective treatment, and following-up early in the disease process, will improve patient outcomes. "The NHS breast cancer screening program was started in 1988 after the Forest Report and is based on mammograms taken every three years to detect changes in breast tissue that may indicate the presence of cancer" [3].

There are several review papers on advanced Machine learning for predicting PCR in breast cancer by analysing MRI data. One of the several studies was done on the Deep Learning Prediction of Pathologic Complete Response in Breast Cancer Using MRI and Other Clinical Data. "The point of this study was to perform a systematic review of deep learning methods using whole-breast MRI images without annotation or tumor segmentation to predict PCR in breast cancer"[4]. In Addition to existing studies, a study was made on a machine learning model that classifies breast cancer Pathologic Complete Response on MRI post-neoadjuvant chemotherapy, "The aim of the following study was to develop and validate a radiomics classifier that classifies breast cancer PCR post-NAC on MRI prior to surgery" [5].

Machine learning methods have the ability to enhance the prediction of PCR and RFS. In this report, we will review the use of machine learning methods for the prediction of PCR and RFS in breast cancer patients. This study will depict the use of both clinically measured features and features derived from MRI prior to chemotherapy treatment. Moreover, it will also review the performance of different machine learning methods and provide recommendations for future research.

2. METHODS

2.1. Dataset

The dataset is a simplified generated dataset based on the public dataset from The American College of Radiology Imaging Network (I-SPY 2 TRIAL). Each patient in this dataset contains 10 clinical features such as Age, ER, PgG, HER2, TrippleNegative Status, Chemotherapy Grade, Tumor Proliferation, Histology Type, Lymph node Status and Tumor Stage, and 107 imaged-based features. These

image-based features were extracted from the tumor region of MRIs using a radiomics feature extraction package. In total, this dataset contains data for 400 patients.

Data instances with null values for classification output were removed from the dataset. It was necessary to remove these, since substituting them would introduce additional error in our training. After removing these, the training set was reduced from 400 instances to 391.

Due to the high dimensionality, the data was trained against three datasets to look for the best performance. For the first, the complete dataset was used, for the second, simple feature reduction techniques were used and for the last, only dimensional reduction through PCA was used.

2.2. Feature selection

Due to the limited data and curse of dimensionality problem, feature selection plays an important role for training the algorithm effectively.
Constant features with variance less than a threshold (0.01) were removed, along with duplicate features. These are some of the minimum requirements for a feature to contain useful information to train our models against.

Features with low correlation against RFS (with a threshold of ±0.05) were removed.
Features with high correlation with another feature (greater than 0.9) were removed.

After feature selection, each feature was normalised into the interval for training.

2.3. Model

After creating and training multiple models with varying hyperparameters, the four main methods tested within this document are SVM, Logistic Regression, Decision Trees, and Neural Networks.

Mean absolute error is used as the loss function to compare the models due to its robustness to outliers.

3. EVALUATION

For each model except the neural networks, grid search is used to exhaustively check through all parameter combinations. This is to tune the hyperparameters for the models to their optimal values.

3.1. Classification

For the classification problem, 4 models are used. These are SVM, logistic regression, decision tree and neural networks.

3.1.1. SVC

For SVM, three grids are used to train the model. The first grid uses a linear kernel, the second grid uses an RBF kernel, and gamma values of [0.5, 0.1, 0.01, 0.001, 0.0001]. The final grid uses a poly kernel, degree values 2-5, and the same gamma values as the second grid. All three grids also use C values of [1, 10, 100, 1000]. After training on both the full dataset and the refined dataset from feature selection, the best results were obtained against the latter and using the RBF kernel, gamma value of 0.01, and the C value of 1. This model obtains an accuracy of 0.84.

3.1.2. Logistic Regression

Two grids are used for logistic regression.; The first uses a liblinear solver, penalty values in ['l1', 'l2'], and C values in [0.001, 0.01, 0.1, 1, 10, 100, 1000]. The second grid uses an L1 loss function, a solver with values ['lbfgs', 'newton-cg', 'sag', 'saga'], and the same C values as the first grid. The model trained with PCA on all datasets using the C value 0.001, an L1 loss function, and a liblinear solver gives the best results, with an accuracy of 0.84.

3.1.3. Decision Tree

With decision trees only one grid is used with a criterion of ['gini', 'entropy']. The splitter of ['best', 'random'] is used to select what strategy to use for the split at each node. A max depth of 10 layers is used. The minimum number of samples to split an internal node is set as 2 - 10. The minimum number of samples for a leaf node is set as 1 - 10 The max features to consider the number of features when looking to best split the data using the following options [None, 'auto', 'sqrt', 'log2']. After training, the best results were found using the complete dataset with the parameters of entropy criterion, a max depth of 10, the max features strategy to be ‘sqrt’ , minimum samples for each leaf being 8, the minimum samples per split to be 9, and using a random splitter. This model achieves an accuracy of 0.84 on the test set.

3.1.4. Neural Network

The neural network was modelled with 2 hidden layers using the Rectified Linear Unit (ReLu) activation function and the Sigmoid activation function for the output layer. When compiling this model,

binary cross entropy is used for the loss and Adam for the optimiser. This model was trained across all three prepared datasets, giving us the best results with PCA. It shows a severe overfitting on the training set by obtaining a result of 99% accuracy, however it only achieves an accuracy of 77% on the testing set.

3.2. Regression

For the regression problem, 3 different models are used. These are SVM, decision tree and neural networks.

3.2.1. SVR

This model uses three grids in its hyperparameter selection process. The first grid has a linear kernel. The second grid uses an RBF kernel and gamma values of [0.5, 0.1, 0.01, 0.001, 0.0001]. The third grid uses a poly kernel with a degree from 2 to 5, and the same gamma values as the second grid. All three grids use C values in the range of [1, 10, 100, 1000]. The best model found is against the feature selection set, and using an RBF kernel, gamma value of 0.5 and C value of 100. This model provides a mean absolute error of 23.65.

3.2.2. Decision Tree

With decision trees only one grid is used with a criterion of ['mse', 'friedman_mse', 'mae']. The splitter of ['best', 'random'] is used to select what strategy to use for the split at each node. A max depth of 10 layers is used. The minimum number of samples to split an internal node is set as 2 - 10. The minimum number of samples for a leaf node is set as 1 - 10 The max features to consider the number of features when looking to best split the data using the following options [None, 'auto', 'sqrt', 'log2']. After training, the best results were obtained from using the complete dataset with the parameters of mse criterion, a max depth of 8, the max features strategy to be ‘log2’, minimum samples for each leaf being 5, the minimum samples per split to be 7, and using a random splitter. This model gives a mean absolute error of 21.27.

3.2.3. Neural Network

The neural network is modelled with 2 hidden layers using the Rectified Linear Unit (ReLu) activation function and the Linear activation function for the output layer. When compiling this model, mean absolute error is used for the loss and Adam for the optimiser. The best result from this method is obtained using the complete dataset. It achieves a mean absolute error on the training set of 14.61, and a mean absolute error of 27.10 on the testing set.

4. DISCUSSION

4.1. Classification

SVM is the best performing classification model that was trained using the dataset provided based on its cross validation and test accuracy.

4.1.1. SVM Advantages

SVM tends to perform well on smaller data sets, and therefore performed well at this task. Regularisation through tuning the C value also helps this model avoid overfitting too much.

4.1.2. SVM Disadvantages

The model would not benefit as greatly from a large increase in the size of the test data. Also, feature selection played a large role in the performance of this model, as it overfit to the features which were providing less useful information for the model.

4.2. Regression

Decision Trees is the best performing regression model that was trained using the dataset provided.

4.2.1. Decision Tree Advantages

Not only does this model perform well, it has a couple of other distinct advantages for easily building a model. This model does not require the data to be normalised to perform well. It also handles null values well in inputs. It therefore requires a lot less preparation than other methods tried to get high levels of accuracy.[6]

4.2.2. Decision Tree Disadvantages

Predictions from a decision tree model are not continuous, which is a limitation for the results of the regression model. Also, a slight change in the initial data may produce a model which behaves very differently for similar inputs. Decision trees can also take a long time to train on large datasets, which may be a problem in other cases.[6]

4.3. Other Methods

The neural network showed severe overfitting and would benefit from a larger dataset to compensate for the size of the dimensional space. [7] It is

incredibly overfitted and would greatly benefit from some form of regularisation. There is good potential for the method to compete with SVM and Decision Trees with more data and good regularisation.

5. CONCLUSIONS

From training multiple models across a varying number of hyperparameters and datasets, we concluded that when it comes to classification, the best performing model against the validation set is SVM, achieving a validation accuracy of 0.84, with a high cross validation accuracy. Decision tree and Logistic regression also performed competitively. For regression, the best performing model against the validation set is from the decision tree method, achieving a mean absolute error of 21.27.

With future work, this type of project can be boosted by having much more data. With the introduction of more data we can create a more thorough neural network to help garner a greater accuracy as with the current level of data it falls short. Some improvements may also be made in additional research through additional regularisation for the models with signs of overfitting.

6. REFERENCES

[1] World Health Organization (WHO). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer (2022).

[2] Hawkes, N. Cancer survival data emphasize importance of early diagnosis. BMJ 364, l408 (2019).

[3] Forrest, P. & Department of Health and Social Security. Great Britain. Breast cancer screening: report to the Health Ministers of England, Wales, Scotland & Northern Ireland. (1986).

[4] Khan, N.; Adam, R.; Huang, P.; Maldjian, T.; Duong, T.Q. Deep Learning Prediction of Pathologic Complete Response in Breast Cancer Using MRI and Other Clinical Data: A Systematic Review. (2022)

[5] Sutton, E.J., Onishi, N., Fehr, D.A. et al. A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy. Breast Cancer Res 22, 57 (2020). ht
tps://doi.org/10.1186/s13058-020-01291-w

[6] Gareth, James; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2015). An Introduction to Statistical Learning New York: Springer- ISBN 978-1-4614-7137-0.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. url: https://www.deeplearningbook.org/