Categorising Breast Cancer as Benign or Malignant Using Machine Learning

Jude Waide

1 Summary

1.1 Introduction

EACH year, 55,920 people in the UK are diagnosed with breast cancer, with 11,499 people dying of breast cancer from 2016-2018 [1]. The earlier a tumor is discovered, the higher the chance of survival. Using machine learning a computer could learn to distinguish between a malignant and benign tumor, potentially helping detect cancers missed by a doctor or help in speeding up the diagnosis process.

1.2 Problem Definition

The goal of this project is to use machine learning to create a model which accurately classifies a breast cancer tumor as malignant or benign based on the cancer's features. This is a supervised binary classification problem as a tumor needs to be placed into 1 of 2 categories and the training data is labelled.

1.3 Proposed Solution

Shown in figure 1, the proposed solution is a bagging ensemble model, the idea being to increase accuracy by reducing variance. The ensemble method uses soft voting and is composed of a decision tree, KNN and Logistic Regression Mode. The decision threshold of the ensemble was then adjusted using a custom method so the final model had a high a recall as possible, without sacrificing precision unnecessarily.

1.4 Results

The results show that the proposed solution performed significantly better than the baseline model, having both a higher recall and accuracy. Changing the decision threshold of the model was successful as it was responsible for a substantial increase in recall without lowering the precision too much. The method used for choosing the decision threshold worked well, given the proposed model had a recall of 0.99, however there is still room for improvement as some malignant tumors were still misdiagnosed.

Potential improvements to the method include; increasing the number of folds in the cross-validation to make finding the desired threshold likelier, or artificially lowering the threshold by a small margin after the original calculation to ensure the recall is 1.

Manuscript received April 19, 2005; revised August 26, 2015.

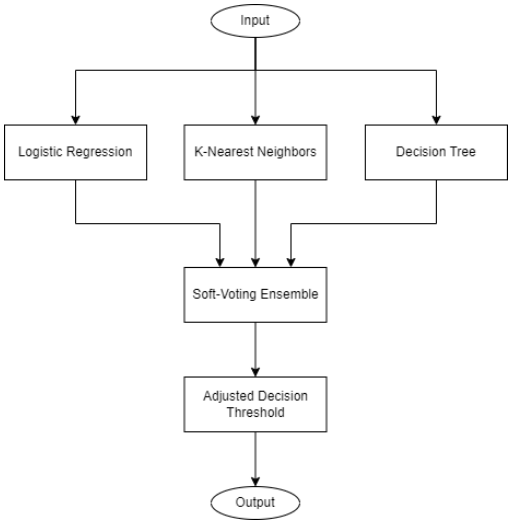

Fig. 1. Diagram of proposed model

Unlike adjusting the decision threshold, the bagging ensemble method used provided no apparent benefits - the increase in accuracy over the baseline was due to the tuning of hyper-parameters and not because an ensemble was used. The KNN model used as part of the ensemble had a higher accuracy than the proposed model before the decision threshold was adjusted, suggesting that it is likely a KNN model with an adjusted threshold would perform better over the proposed model. It is also possible that a different ensemble method such as boosting or stacking may provide better results.

2 Data and Experimental Setup

2.1 Data-set analysis and manipulation

THE Wisconsin Breast Cancer dataset [2] provides 699 instances of tumors, each with 9 key features on a scale of 1-10. There are no missing values except for the 'Bare nuclei' column, in which there are 16 missing values. These missing values were filled with the column average so these rows could still be used without adding much bias. More information on what these features are and why they're relevant can be found at [3].

Figure 2 shows this data-set is already clean with all features occupying the same range with a similar mean, so no scaling was performed.

Feature	Count	Mean	std
Clump thickness	699	4.42	2.82
Uniformity of cell size	699	3.13	3.05
Uniformity of cell shape	699	3.21	2.97
Marginal adhesion	699	2.81	2.86
Single epithelial cell size	699	3.22	2.21
Bare nuclei	683	3.54	3.64
Bland Chromatin	699	3.44	2.44
Normal Nucleoli	699	2.87	3.05
Mitoses	699	1.59	1.72

TABLE 1
Table displaying characteristics of data-set features

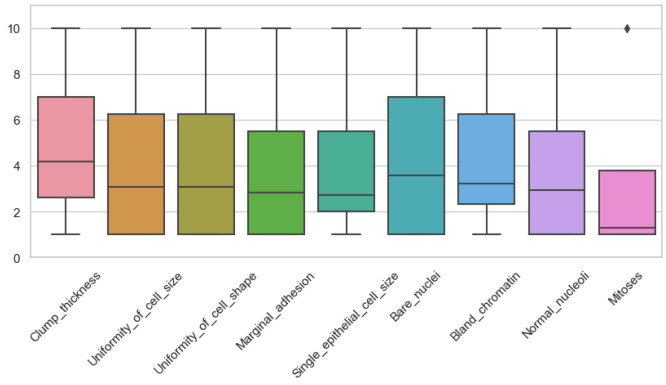

Fig. 2. Boxplot of features, showing range and mean

The target variable is the class of tumor, 2 signifying benign and 4 signifying malignant. This was remapped so that benign is 0 and malignant is 1 as these values are more convenient when working with binary classifiers in Sklearn.

Figure 3 shows that all of the 9 features are positively correlated with the target variable, with 8 out of the 9 being strongly correlated. Therefore no features were dropped as all are relevant in predicting whether a tumor is malignant.

2.2 Data-set split

The data-set was split into a training set and testing set with a 60/40 ratio respectively. The proportion of malignant tumors to non-malignant tumors was kept the same in both sets. As a general rule of thumb, there needs to be at least

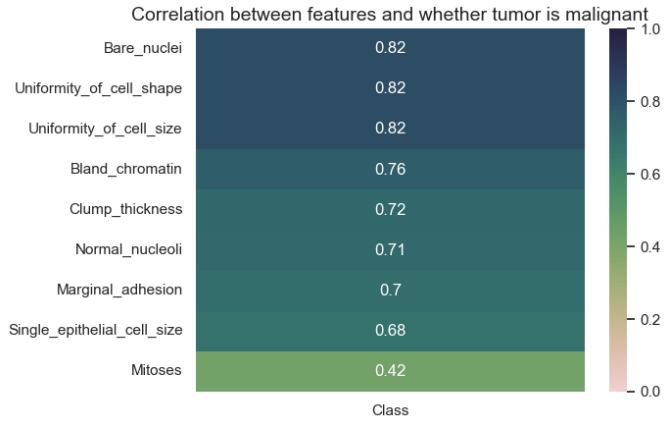

Fig. 3. Heatmap showing the correlation between each feature and the target variable

10 times the number of data points as there are features to effectively train a model [4]. With this split there will be 419 data points for training, which is much more than the required 90, whilst still leaving plenty of data points for testing the model.

2.3 Evaluation Metrics

To evaluate the model, the most important metric in this case is recall. A false negative (identifying a tumor as benign when it is actually malignant) is far more dangerous than a false positive (the inverse). This is because if a malignant tumor is missed, it may only be diagnosed at a later date and a patient's survival rate decreases the longer a tumor is left undetected [5]. On the other hand, a false positive is not dangerous to a patient as any tumors flagged as malignant would be looked at by a doctor anyway.

The second most important metric is accuracy. If a model always identifies a tumor as malignant, it would have a high recall but wouldn't be useful in any way. For these reasons, a model should never diagnose a malignant tumor as benign, and second to this identify as few benign tumors as malignant as possible.

3 Model Evaluation

3.1 Model Selection

The proposed model to solve this solution is an ensemble bagging method combining a binary tree, logistic regression and k-nearest neighbours model using soft voting with an adjusted decision threshold. For a baseline model to compare this to, a decision tree was chosen as it is relatively simple model suited for binary classification tasks.

The proposed solution should be better than the baseline model in theory for two reasons. Different models have different strengths and weaknesses, so by combing different models using bagging the variance should be reduced over the baseline model. This is because the impact of individual weaknesses are compensated for by the other models, resulting in an overall higher accuracy.

The second reason is that by adjusting the decision threshold, precision can be exchanged for recall - the most important metric for this problem. This should mean the proposed model diagnoses far fewer malignant tumors as benign compared to the baseline model.

3.2 Training Process

The baseline model was trained using Sklearn's DecisionTreeClassifier, with default parameters and hyperparameters. The three models in the ensemble were trained using Sklearn's DecisionTreeClassifier, KNeighborsClassifier, LogisticRegression using default parameters. The hyperparameters for the three models were tuned using a 5-fold cross-validation grid search.

The 5-fold cross-validation splits the training set into 5 folds, rotating one to be used as a validation set. If the hyperparameters were tested on the testing-set during tuning, the final full-model test would not be an accurate representation of real-world performance. If they were trained and tested using the full training set, they would be overfit to the training set and so the model would perform worse. By

using the k-fold validation however, both of these issues are overcome. The exact parameters and values checked can be seen in table 2, with any hyper-parameters not listed left as the default.

Model	Hyper-parameter	Values	Chosen value
LogReg	C	Powers of 10 from 0.01 to 100	0.1
	solver	newton-cg, lbfgs, liblinear	newton-cg
Knn	n_neighbors	Integers from 1 to 15	9
	metric	euclidean, manhatten, minkowski	euclidean
DecisionTree	criterion	gini, entropy, log_loss	gini
	max_depth	1,3,5,7,10,15	5
	min_samples_leaf	1,3,5,10,15,20	3
	min_samples_split	even integers from 6 to 18	6

TABLE 2
Hyper parameters chosen for models

After hyper-parameter tuning, the three models were passed to a VotingClassifier using soft voting. Soft voting (summing probabilities following by a threshold) was used as there is no decision threshold which can be adjusted when using hard voting (majority vote).

A new decision threshold was then found using a form of 5-fold cross-validation. For each fold, a decision threshold was found by decreasing the threshold from 1 until the recall was 1 when tested on the validation set. The minimum of these 5 was then selected as the final decision threshold, the hope being to ensure a recall of 1 with the highest accuracy possible when tested using the testing set. The threshold was found in this manner, as if the threshold was adjusted based on training set then the model would be overfit to the training-set so an accurate measure of how the model would perform in the real world would not be possible. The final threshold obtained in this instance was 0.231.

3.3 Performance Comparison

Table 3 shows that overall the proposed model was better than the baseline model - it's recall of 0.990 is a significant improvement compared to the baseline of 0.907. The recall of the ensemble before the decision threshold was adjusted was 0.969, showing that adjusting the decision threshold was successful in improving recall.

Model	Accuracy	Recall	Precision
Baseline	0.957	0.907	0.967
LogReg	0.975	0.959	0.969
KNN	0.979	0.969	0.969
Tree	0.961	0.979	0.913
Ensemble	0.975	0.969	0.959
Proposed Model	0.968	0.990	0.923

TABLE 3
Table showing various performance metrics of the models. LogReg, KNN and Tree are the 3 models used in the ensemble of the proposed method. 'Ensemble' is the proposed method before the threshold was adjusted.

Figure 4 demonstrates this high recall, with the proposed model only mislabelling 1 out of the 97 malignant tumors in the test set, compared to the baseline model mislabelling 9. This indicates the method by which the decision threshold was selected was successful, but given that the recall is not 1, there is still room for improvement.

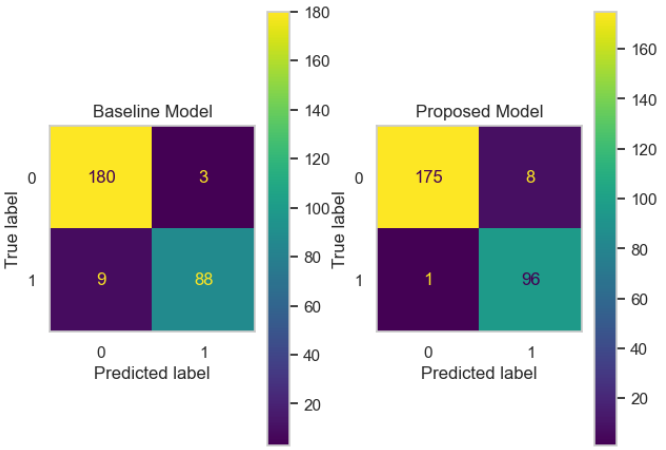

Fig. 4. Confusion matrix of baseline and proposed model. 1 is malignant, 0 is is benign.

Not only is the recall of the proposed model higher than the baseline, but so is the overall accuracy with the proposed model having an accuracy of 0.968 compared to the baseline 0.957. However, the results in figure 3 indicate that this increase in accuracy cannot be attributed to the use of an ensemble method, as there is no improvement to the accuracy (before threshold is adjusted) compared to the models which make up the ensemble. This suggests that the improvement instead comes from the tuning of hyper-parameters and the higher accuracy of LogReg and KNN compared to a decision tree.

As KNN has a higher accuracy than the ensemble and the other two models used in the ensemble, it is possible that a KNN model with an adjusted threshold would perform better than the proposed model and that an ensemble is unnecessary.

4 Self Evaluation

4.1 What have I learnt from the lectures?

From the lectures I have gained an understanding about the different types of machine learning problems, e.g. regression vs classification and supervised vs unsupervised. I've also learnt which models are suited to which tasks as well as the maths behind some of the simpler models.

As well as this, I'm now aware of some of the common problems which you can encounter when training a machine learning model (i.e. underfitting and overfitting), and why you should split data into training, test and validation sets to avoid these.

4.2 What have I learnt from the coursework?

Through the process of making this coursework I've gained experience at using LaTeX, as well as scientific analysis, which I think will be very useful when I hopefully write academic papers in the future. I've also got practice at using sklearn, so I now feel confident in training my own machine learning models.

4.3 What were my difficulties in the module

I sometimes struggled to follow the maths shown in lectures and I often had to look at it again afterwards before I fully understood what was happening, although even then I sometimes found the lecture slides somewhat challenging to follow. Asides from this I haven't had any difficulties with the module and I think I've understood the majority of the content.

4.4 What would I do differently if I were to do it again?

I struggle to think of much I would do differently. I've attended all of the lectures and have been engaged with the material so I think I've made the most out of what the module has had to offer me and have enjoyed it overall. If there's one thing I would do differently though I think it would be to read more literature about what we were taught as this is an area which of great interest to me, however there's plenty of time for that going forward.

4.5 Unique contributions or novel ideas

I don't know what has and what hasn't been done before and I'd be surprised if I have done anything novel, however the most likely candidate is the particular way in which I used cross validation to find a decision threshold which maximises recall without sacrificing accuracy unnecessarily.

REFERENCES

[1] Cancer Research UK, "Breast cancer statistics," *Cancer Research UK*, Nov. 13, 2022. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer (accessed Jan. 08, 2024).
[2] Wolberg,WIlliam. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.
[3] OpenNN, "Classification Example: Breast cancer diagnosis | Tutorial," *www.opennn.net*. https://www.opennn.net/documentation/classification_example.html (accessed Jan. 08, 2024).
[4] "One in ten rule," *Wikipedia*, Jan. 30, 2023. https://en.wikipedia.org/wiki/One_in_ten_rule (accessed Jan. 08, 2024).
[5] Cancer Research UK, "Survival for breast cancer | Breast Cancer | Cancer Research UK," *www.cancerresearchuk.org*, Jan. 03, 2020. https://www.cancerresearchuk.org/about-cancer/breast-cancer/survival (accessed Jan. 08, 2024).