A Comparison of Decision Tree and Random Forest Algorithms on Breast Cancer Prediction

Tugba Sanver

Description and Motivation of the Problem

The aim of this research is comparing performances of two Machine Learning algorithms which are Decision Tree and Random Forest on predicting classification labels of breast cancer cell attributes and measurements dataset to help to detect potential cancer cases. This problem is binary classification problem regarding there are only two target labels which are ‘malignant’ and ‘benign’ for researched cells, therefore it will be solved via supervised learning algorithms. A previous study from Yixuan Li and Zixuan Chen(2018) is considered to compare methodologies and results on the same dataset(1).

Initial Analysis of the Dataset

- The dataset is Breast Cancer Wisconsin(Original) from UCI Machine Learning Repository. [8]
- The dataset contains 699 samples and 10 columns which are 9 predictive features and 1 target variable. All variables are numeric. The target variable has two labels which are ‘2’ for benign labelled cells and ‘4’ for malignant ones.
- Only one column(bare nucleoli) had missing values, it is imputed with mean value of the column.
- The basic statistics of all variables is shown on table 1.
- Heatmap visualisation is used to analyse the correlations of variables, as seen in figure 2. It can be considered as size uniformity, shape uniformity and bare nucleoli attributes are the factors which mostly affect cell diagnosis.
- Pie chart in figure 1 indicates there is imbalance between label classes. This may cause an imbalanced classification because there are too few examples of the minority class to learn the decision boundary properly. Oversampling the minority class examples can solve this issue. In this purpose, class labels are balanced using SMOTE(Synthetic Minority Oversampling Technique) approach[2]. The possible problem is eliminated with this solution.
- Figure 4 shows the distributions of the variables according to label classes via histogram. High values of bare nucleoli, clump thickness and bland chromatin usually indicates malignant diagnosis.

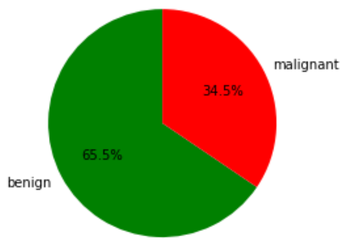

Figure 1. Class label distribution

	count	mean	std	min	25%	50%	75%	max
clump_thickness	699.0	4.417740	2.815741	1.0	2.0	4.0	6.0	10.0
size_uniformity	699.0	3.134478	3.051459	1.0	1.0	1.0	5.0	10.0
shape_uniformity	699.0	3.207439	2.971913	1.0	1.0	1.0	5.0	10.0
marginal_adhesion	699.0	2.806867	2.855379	1.0	1.0	1.0	4.0	10.0
epithelial_size	699.0	3.216023	2.214300	1.0	2.0	2.0	4.0	10.0
bare_nucleoli	699.0	3.532189	3.602774	1.0	1.0	1.0	5.0	10.0
bland_chromatin	699.0	3.437768	2.438364	1.0	2.0	3.0	5.0	10.0
normal_nucleoli	699.0	2.866953	3.053634	1.0	1.0	1.0	4.0	10.0
mitoses	699.0	1.589413	1.715078	1.0	1.0	1.0	1.0	10.0
diagnosis	699.0	2.689557	0.951273	2.0	2.0	2.0	4.0	4.0

Table 1. Basis statistics for the data set

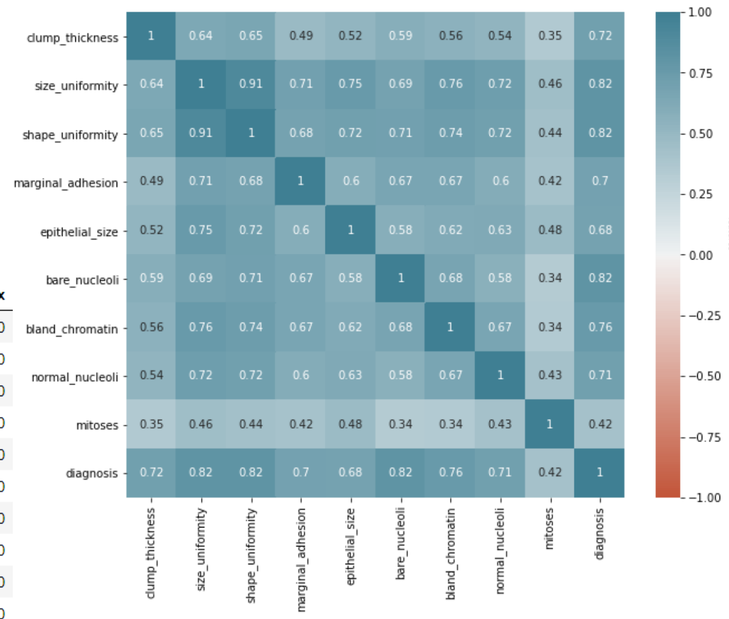

Figure 2.Heatmap visualisation for correlations of all variables

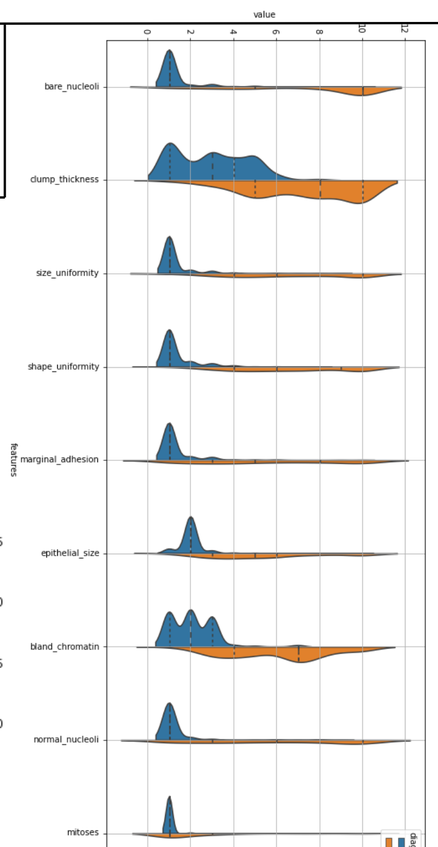

Figure 4. Violin plot for distributions through labels

Descriptions of Algorithms, Advantages and Disadvantages

Decision Tree

- Decision tree is Classification and regression trees(CART) model which is a supervised learning algorithm which contains recursive partitioning of input place to solve classification and regression problems.
- It usually represented by a tree and its leaves and has a flowchart-like structure.
- The data are split according to the feature values. Each internal node shows a test on a feature, each branch shows the result of the test, and each leaf node(a node that has no children) shows a class label[3].

Advantages:

- They are easy to interpret and fast to produce.[5]
- They can handle mixed discrete and continuous inputs easily.
- They make automatic feature selection.

Disadvantages:

- They have lower accuracy levels comparing other models, especially for the bigger data sets.
- They are unstable in terms of small changes made to the input data, it can trigger large effects on prediction. That causes high variance for the model.[6]

Random Forest

- A decision tree based supervised learning algorithm which uses a random subset of the training samples for each tree.
- It also uses a random subset of features in each step of growing each tree.
- It averages their outputs for any given input x via a technique called as Bagging(Bootstrap Aggregating).[7]

Advantages:

- It lowers the model variance, therefore Random Forest models are more resistant to overfitting.[6]
- Random forest lead the robust models which have good prediction accuracy, especially for the bigger data sets[1].
- It eases to measure the relative importance of each feature on the prediction.[7]

Disadvantages:

- Training part for Random Forest model is fairly slow comparing many other models, especially with complex data with many features .[6]
- It is more difficult to interpret than a decision tree model.

Hypothesis statement

- Random Forest algorithm is able to produce better accuracy and f1 scores than Decision Tree algorithms according to similar breast cancer classification researches[1], especially for the large datasets[3].
- Random Forest with hyperparameter optimisation has high accuracy performance on models.[4]
- Decision Tree algorithm is more efficient on running time and memory size than Random Forest.[3]

Description of the choice of training and evaluation methodology

- The original dataset was split into training and testing sets, 80% and 20% respectively. Training dataset has 559 samples, and the test dataset has 140 samples.
- Cross validation with 10 folds is used on training dataset to calculate classification error for hyperparameter optimisation models.
- Best feasible hyperparameter optimisation values are selected via Bayesian optimisation search and grid search for both models. Also Bayesian optimisation and grid search performances are compared.
- Feature selection approach is applied to simple Random Forest algorithm as an experiment to see its effectiveness on classification error.
- Classification errors and validation accuracies are compared between simple models and optimised models for each algorithm. The models have less validation losses and higher accuracy scores are selected.
- As an experiment and a possible scenario, a grid search for the best ensemble algorithm with hyperparameter optimisation is trained.
- Best models for each algorithm is fitted to the test dataset. Accuracy, recall, precision, f1 score, AUC scores and ROC curves of each model prediction are calculated and compared between algorithms.

Choice of parameters and experimental results

Decision Tree

- Maximum number of splits and minimum leaf size hyperparameters are used in optimal hyperparameter searches via grid search. Classification error is 0.0322. Out of bag cross validation losses(Observed objective function values) is 0.0429.
- Maximum number of objective function evaluation for grid search optimisation is limited at 30 since there is no further validation error improvement beyond that(figure 5).
- The decision tree model with grid search hyperparameter optimisation picked the 3 minimum leaf size and 8 Maximum splits.
- Validation accuracy score for the decision tree model with grid search hyperparameter optimisation is 0.97. AUC score for this is 0.98.

Random Forest

- Hyperparameter optimisation with number of learning cycles, minimum leaf size, learn rate and maximum number of splits is used in a grid searched model to find optimal values of hyperparameters to improve model. The grid search picked the values that 210 for learning cycles, 0.046 for learning rate, 12 for minimum leaf size and 8 maximum splits.
- Classification error for training is 0.0161.
- Maximum number of objective function evaluation for grid search optimisation is limited at 30 since there is no further validation error improvement beyond that(figure 5).
- Validation accuracy score for the decision tree model with grid search hyperparameter optimisation is 0.98. AUC score for this is 0.999.

Comparison

- Final models are fitted on the unseen data and predictions are made. AUC, accuracy, f1, precision and recall scores of both model prediction are compared in figure 6. For Decision Tree and Random Forest models, AUC scores are 0.96 and 0.99, accuracy scores are 0.94 and 0.97, f1 scores are 0.95 and 0.98, precision scores are 0.98 and 0.99, recall scores are 0.93 and 0.97, respectively.
- ROC curve comparison of the chosen models can be seen in figure 7.

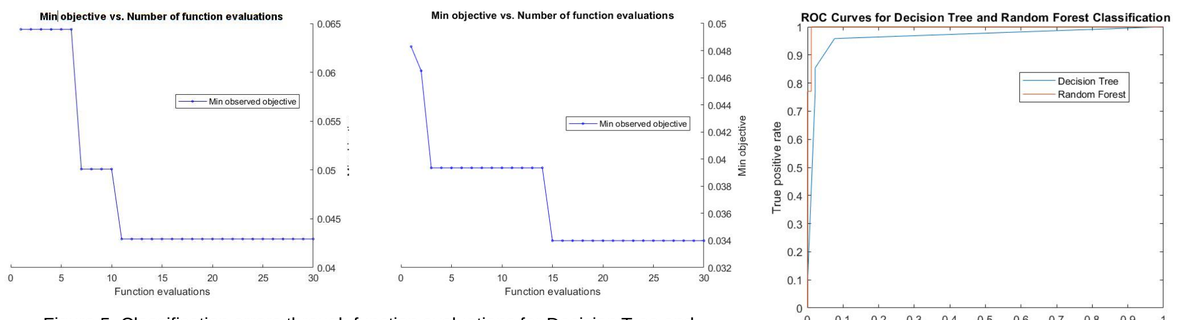

Figure 5. Classification errors through function evaluations for Decision Tree and Random Forest models on training

Figure 6. ROC curve comparison of the models on the test data

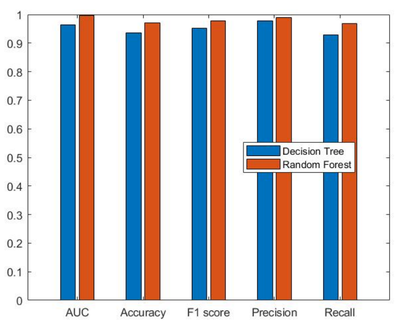

Figure 7. Metric comparison of the results of the models on unseen data

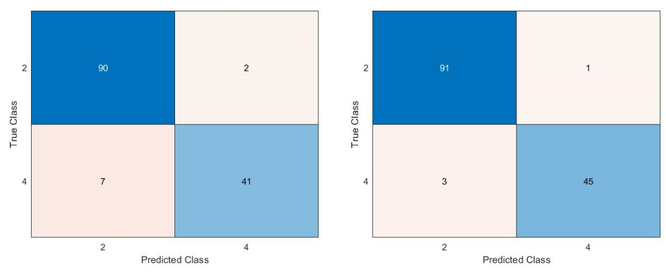

Figure 8. Confusion matrixes of Decision Tree and Random Forest models on the test data

Analysis and critical evaluation of results

- Evaluation metrics(AUC, accuracy, f1, prediction and recall) indicates Random Forest model has higher performance than Decision Tree model for predicting unseen test data. Difference between metrics are changing between 0.1 to 0.4, which are not hugely significant but considered as important. If the data set were smaller, the metric could be same, or even Decision Tree model could be better according to the previous studies.[3] Since the data set balanced after SMOTE, the accuracy metric would be enough to compare the models, however it is considered as analysing all metrics would be better.
- Training time was only 5 seconds for Decision Tree model, comparing 150 seconds of Random Forest model’s training time it trains significantly faster than Random Forest algorithm, which is one of the most important differences for training performance costs between models. Testing times were same for both models, so it cannot be concluded about prediction time efficiency comparison for testing.
- 10 folds cross validation is performed for the Random Forest model with hyperparameter optimisation to achieve fair comparison of models since tuned Decision Tree model has 10 folds cross validation also, although Random Forest algorithm consists of different bags like folds of cross validation systematic and has out of error like cross validation error.
- As Random Forest algorithm uses longer time to train the data set with many features as proposed as a hypothesis statement, feature selection is applied to simple Random Forest model as an experiment. It did not improve classification error of the model on the training data with 9 predictor features.
- ROC Curve comparison of the model predictions indicates Random Forest algorithm predicted almost all labels true as its true positive rate is almost 1 and false positive rate is almost 0. Also, confusion matrix(figure 8, second matrix) of Random Tree supports this argument as just 1 false positive and 3 false negative counts for 140 samples on the test data.
- Maximum numbers of objective function evaluation for grid searched hyperparameter optimisations did not show any improvement on validation losses after 15 evaluations, since the maximum number is set for 30, to demonstrate the graph better.
- Grid searches for both models picked the same maximum split number which is 8. Similar research paper shown the same split number for hyperparameter tuning on same data set using Random Forest. Some other hyperparameters are not consistent, therefore there can be more than one best feasible options for some hyperparameter values, although some other can have less options in terms of optimisation.
- Decision Tree model has 0.97 validation accuracy score and 0.94 test accuracy score, on the other hand Random Forest model has 0.98 validation accuracy score and 0.97 test accuracy score. It can be concluded that regarding similarity of accuracy scores of validation and testing, Random Forest is resistant to overfitting more, as expected.

Lessons Learned and Future Work:

Lessons Learned

- Maximum numbers of objective function evaluation should be lower if any improvement does not occur for validation error beyond a specific point.
- Feature selection does not always help the model performance, especially for the data set consist of 9 predictors.
- Limitation of number of grown trees can help for a faster training where classification error is not improve any more.
- Validation errors are higher than normal training errors, and more realistic to closeness to test errors.

Future Work

- Pruning can be considered to decrease variance of the decision tree model.
- Outlier detection and removal for the original data set may help improvement on prediction quality of the decision tree model.
- Different feature selection approaches may be planned to perform onto Random Forest model.

References

1. Yixuan Li, Zixuan Chen. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. Applied and Computational Mathematics. Vol. 7, No. 4, 2018, pp. 212-216. doi: 10.11648/j.acm.20180704.15
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
3. Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.
4. Simon Bernard, Laurent Heutte, Sébastien Adam. Influence of Hyperparameters on Random Forest Accuracy. International Workshop on Multiple Classifier Systems (MCS), Jun 2009, Reykjavik, Iceland. pp.171-180, ff10.1007/978-3-642-02326-2_18ff.
5. Quinlan, J. Learning Logical Definitions from Relations. Machine Learning 5, 239–266 (1990). https://doi.org/10.1023/A:1022699322624
6. Murphy, K. P. (2012). Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press.
7. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
8. Wolberg, William. (1992) Breast Cancer Wisconsin (Original). UCI Machine Learning Repository.
9. Buttan Y., Chaudhary A., Saxena K. (2021) An Improved Model for Breast Cancer Classification Using Random Forest with Grid Search Method. In: Goyal D., Chaturvedi P., Nagar A.K., Purohit S.