A comparison of Naïve Bayes and Random Forest algorithms for predicting recurrence events of breast cancer

Ruth Wetters 200000753

Brief description and motivation of the problem (5%)

Breast cancer is the most common cancer affecting women and the second most common overall[1]. Machine learning is increasingly used in the field of biomedical research to diagnose complex illnesses like cancer, which can aid early interventions, reduce treatment costs, and improve patient outcomes.

This study will compare and contrast the standard and modified performance of Naive Bayes and Random Forest in a binomial classification problem based on the breast cancer recurrence dataset available from the UCI Machine Learning repository, with reference to a 2018 paper by Bharati et al.[1]. Naive Bayes is often considered a good baseline model, whereas Random Forest is an ensemble method suited for more complex analysis, so comparing these two should allow plenty of scope for analysis.

Initial assessment of the dataset including basic statistics (5%)

Description: UCI ML Breast Cancer dataset

Instances: 285

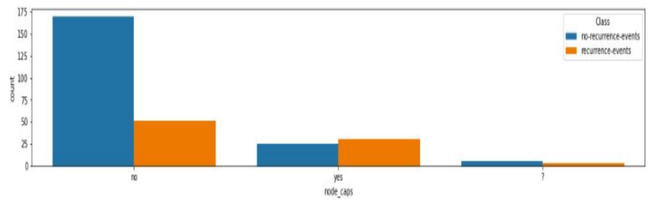
The target has two classes: tumour classed as malignant or benign (originally represented as 'M' and 'B' but I have substituted 1 and 0). This dataset is slightly imbalanced (65% versus 35%) but I did not consider it a class imbalance problem as there is still significant representation of both classes. However, it should be noted that any results below 65% are no better than random chance.

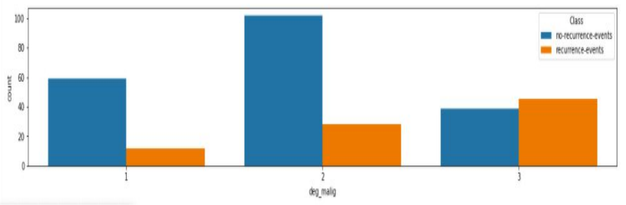
Attributes: 10 categorical variables, of which three were originally were numeric but have been binned by original researchers

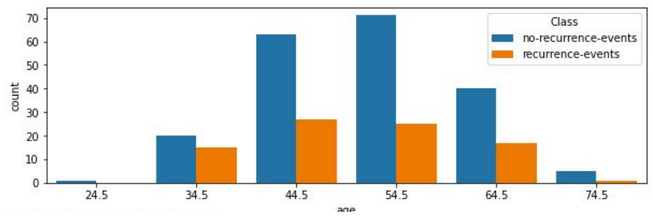
Missing data: yes, in 8 instances one feature is missing, however I did not remove those instances as both Naive Bayes and Random Forest are capable of handling missing values.

Basic statistics

- By using the average value for numeric variables age, inv_node and tumour_size, can calculate approximate summary statistics
- Mean tumour size was 29cm in recurrence events and 25cm in non-recurrence; standard deviation was 8.7 for recurrence compared to 11 for no recurrence
- Bar charts show that age, tumour size and inv. nodes follow a roughly normal distribution for both target classes
- Tumour size, inv. Nodes, node-caps, all show correlation between higher values and increased probability of recurrence events.
- Weak correlation between age and menopause (weak because of the binning).







Brief summary of the two ML models with their pros and cons

Naive Bayes (NB)

- NB is a comparatively simple algorithm commonly used for classification problems such as text classification and medical diagnosis
- It consists of a simple Bayesian network, with no assumption of correlation between variables
- NB is a generative method, meaning that the posterior is not directly modelled but calculated using the joint probability.
- The goal of Naive Bayes is to calculate conditional probability for each of k possible outcomes Ck. Maximum likelihood estimate (MLE) is used to estimate parameters: prior probability and conditional probability. The prior probability equals the number of certain cases of y occurrence divided by total cases. The Naïve Bayes classifier combines the Bayes probability model with a decision rule, most often the maximum a probability (MAP) rule.

Pros:

- can be trained easily and quickly and can be used as a benchmark model
- Works well on high-dimensional data
- Requires a small amount of data to converge
- Can make probabilistic predictions

Cons:

- Can be affected by outliers
- Assumption of independence is flawed and can reduce performance

Random Forest (RF)

- RF is an ensemble approach to machine learning based on the Decision Tree algorithm and pioneered by Leo Breiman[2].
- Builds an ensemble of decision trees, with each tree splitting at a randomly chosen node from the dataset
- Each point is tested on each tree, and the algorithm makes its prediction based on the majority result
- RF is an example of a discriminative model, which estimates the posterior then combines with a decision rule.
- It is frequently used in large, high dimensional datasets for its good overall performance and resistance to overfitting.

Pros:

- One of the most effective methods for dealing with large datasets
- Use of multiple trees means this method naturally cross-validates which reduces overfitting
- Can be used to rank features according to their importance
- Easily interpretable
- Can handle both categorical and continuous variables
- Can handle missing values

Cons:

- One of the more computationally expensive methods
- Does not perform well with sparse data
- Can sometimes have higher bias than a single decision tree

Hypothesis statement

- Based on the results obtained by Bharati et al., I expect Naive Bayes to outperform Random Forest on accuracy, precision, recall and area under the curve.
- Bharati et al. reported 71.6783% accuracy, for Naive Bayes, compared to 69.5804% for RF. NB also scored higher on precision, recall, F-measure, ROC area, and MCC.
- I also expect NB to have a lower training time, due to the simple nature of the model
- I expect that feature engineering will have little effect on model performance, but that hyperparameter tuning will improve both

Description of choice of training and evaluation methodology (5%)

- I will show how feature selection can be performed in order to improve models
- Evaluate the impact of hyperparameter tuning on each model
- Compare the decision boundary for NB and RF
- Will cross-validate each model
- I will evaluate models on their accuracy, AUC and training time
- Will evaluate performance using a 70% train, 30% test split

Choice of parameters and experimental results

Naive Bayes

Parameters

- Conducted feature selection and ran the model with + without less useful variables
- Manipulated the prior to measure effect on accuracy, recall, precision and AUC

Results

- Feature selection identified the most important variables as tumour size, node-caps and breast-quadrant.
- Manipulating the prior significantly improved accuracy
- Approximating numeric averages for pre-binned numeric variables reduced the accuracy

Random Forest

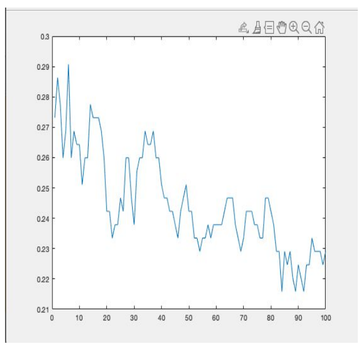
Parameters

- Varied number of trees
- Varied number of splits
- No variables removed

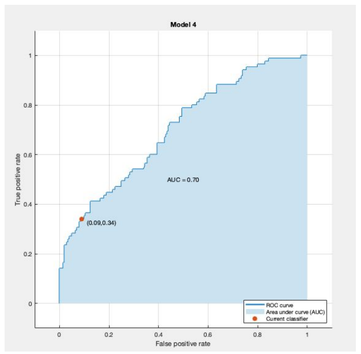
Results

- A large number of trees did not improve the model: the ideal number was found to be 100
- Best number of splits was found to be 38

	Naive Bayes	Random Forest
Accuracy	0.7956	0.6400
AUC	0.6255	0.9605
Training time	1.7132	3.8590



Number of splits in RF compared to generalisation error



A bagged decision tree sample from the Classification Learner

Analysis and critical evaluation of results (20%)

- This is an interesting result: the NB model performs much better on accuracy and trains quicker, but the RF model has an extremely high AUC. AUC is generally considered a broader metric, as it is an aggregate measure across classification categories.
- This result could be due to the class imbalance problem, or it could be that NB has high bias, and low variance.
- RF is generally considered a model which does not have especially high bias or variance and does not overfit, according to Breiman[2].
- I would consider NB the superior model in this case, as in medical diagnosis, high accuracy is more likely to render the model useable.
- In this case the NB assumption of independence held, in line with the results of Bharati et al.

- The NB model was extremely responsive to the tuning of the prior, which is reflective of the fact that NB is much more dependent on manual optimisation compared to RF. This can be good to maximise performance of a small, specific dataset, but would make it much more difficult to use on a large dataset/number of different datasets.
- I believe that with accurate numeric variables, a kernel distribution would further improve accuracy, and provide scope for optimisation through kernel width. This would also provide scope to tune the number of bins used by the model.
- The RF model was resistant to hyperparameter tuning, which could be because of the sparseness of the data: a better way to improve it could be through bootstrapping, or using larger datasets. Alternatively, because the features were only categorical, there were no attributes that could be binned, and it limited the scope for statistical methods.
- This dataset has no problems with the 'curse of dimensionality', but is nonetheless extremely limited by its size.
- Random Forest is considered self-cross-validating, so it was perhaps unnecessary to use crossval, but I implemented it to enable comparison with NB. I do not think that cross validation had much effect on the RF model, but it is likely to have improved the NB model, to reduce bias and generalisation error.

References

[1] S. Bharati, M. A. Rahman, and P. Podder, 'Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA', in 2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT), Sep. 2018, pp. 581–584, doi: 10.1109/CEEICT.2018.8628084.
[2] L. Breiman, 'Random Forests', Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[3] V. Chaurasia and S. Pal, 'Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer', SN Computer Science, vol. 1, no. 5, Sep. 2020, doi: 10.1007/s42979-020-00296-8.
[4] G. Ciaburro, MATLAB for Machine Learning. Packt Publishing Ltd, 2017.
[5] E. C. Garrido-Merchán and D. Hernández-Lobato, 'Dealing with categorical and integer-valued variables in Bayesian Optimization with Gaussian processes', Neurocomputing, vol. 380, pp. 20–35, Mar. 2020, doi: 10.1016/j.neucom.2019.11.004.
[6] P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri, 'On Supervised Selection of Bayesian Networks', arXiv:1301.6710 [cs, stat], Jan. 2013, Accessed: Dec. 13, 2020. [Online]. Available: http://arxiv.org/abs/1301.6710.
[7] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, 'Data Preprocessing for Supervised Learning', International Journal of Computer Science, vol. 1, pp. 111–117, Jan. 2006.
[8] A. Singh, M. N., and R. Lakshmiganthan, 'Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms', International Journal of Advanced Computer Science and Applications, vol. 8, no. 12, 2017, doi: 10.14569/IJACSA.2017.081201.
[9] 'Using machine learning techniques to predict the recurrence of breast cancer | LinkedIn'. https://www.linkedin.com/pulse/using-machine-learning-techniques-predict-recurrence-breast-alva/ (accessed Dec. 13, 2020).

Lessons learned and future work

- Although RF generally performs better than NB on most datasets, in this case it would be computationally more sensible to opt for a simpler classifier such as Decision Tree which might perform better on a small dataset
- Feature engineering and hyperparameter tuning can improve both RF and NB
- Potential to repeat the methodology with age, tumour size and inv. nodes as continuous numeric variables which would improve precision. Number of bins could then be a hyperparameter for NB
- The two models could be combined to create a Bayesian Random Forest which would theoretically be more accurate
- Could use techniques such as S.M.O.T.E to correct class imbalance by generating new data
- Bharati et al. also compared logistic regression, k-nearest neighbours, and multilayer perceptron - this research could be replicated and analysed

CITY UNIVERSITY OF LONDON EST 1894