Supplementary Material

Glossary

Bayesian network: a probabilistic graphic model, which uses Bayesian inference to calculate probabilities. Bayesian networks consist of roots, leaves and nodes. Bayesian inference is, at its core, the application of Bayes' Theorem: P(A|B) = P(A) P(B|A)/P(B). Naive Bayes is an example of a simple Bayesian network

Maximum-relevance-minimum-redundancy (mRMR): a common approach to feature selection which can identify relevant features in a high-dimensional dataset. It has been found to provide a good balance between relevance and redundancy.

Pearsons/Chi-squared feature selection: a statistical hypothesis that assumes a null hypothesis to calculate a chi-squared distribution. Although its basic use is to prove or disprove the null hypothesis with respect to each variable, it can also be used to rank variables by relevance.

Accuracy: Accuracy is a measure of statistical bias, and is calculated as the number of correct responses as a proportion of total responses.

From reference paper

Kappa statistic: a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. By taking into account random chance, it is a better measurement than accuracy.

MCC: Matthew's Correlation Coefficient helps with the selection of a confusion matrix, in circumstances where there may be multiple diverging matrices. Essentially, it is a correlation coefficient. It profits values between -1 and +1.

Domain-specific terminology

Inv-nodes: the number of axillary lymph nodes containing metastatic breast cancer which can be seen from histological examination.

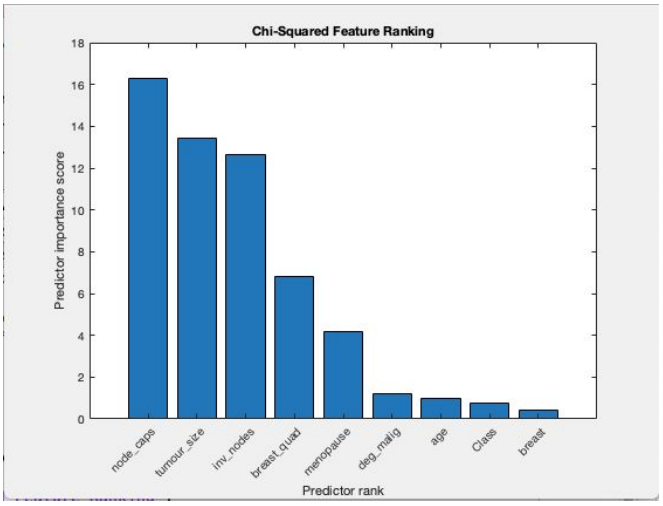
Node caps: "if the cancer does metastasise to a lymph node, although outside the original site of the tumor it may remain "contained" by the capsule of the lymph node. However, over time, and with more aggressive disease, the tumor may replace the lymph node and then penetrate the capsule, allowing it to invade the surrounding tissues"[1].

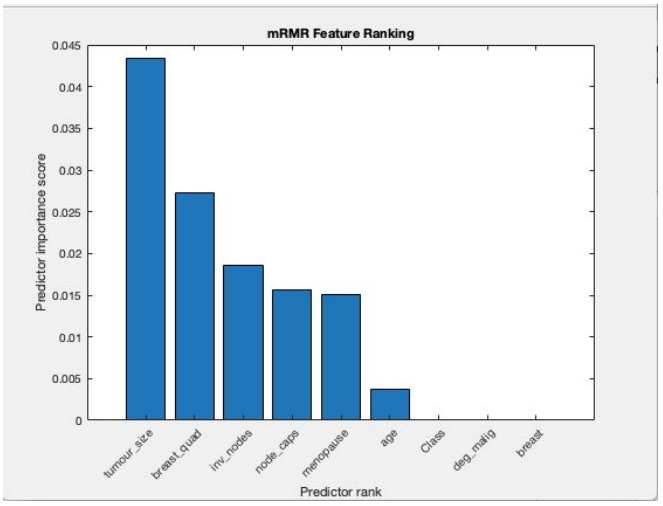
[1] Bharati, "Breast Cancer Prediction", 2018.

Supplementary Material

Relevant intermediate results

Results of feature ranking





There is significant variation between these two ranking algorithms, which suggests that there may be variation in the data. The relatively strong performance of ‘breast_quad’ is due to a sampling bias in the original dataset, according to

Sample of measures taken to improve models

These steps were taken and then cancelled, so are not iterative, as some reduced the accuracy and the ones that improved it, such as predictor importance, increased the training time.

Random Forest

Measure taken	Accuracy
Initial model	0.6850
Num trees = 300	0.6700
Predictor importance on	0.7000
Removed ‘breast’ and ‘deg-malign’	0.6600

Naive Bayes

Measure taken	Accuracy
Initial model	0.7200
Prior = [0.7 0.3]	0.7246
Prior = [0.8 0.2]	0.7956

Supplementary Material

Dummy variable mean threshold on numeric variables	0.6500

Implementation details

General

As a way of representing the thresholded numeric values (age, inv-nodes and tumour-size), I also tested a version of the models with these bins represented by their averages, eg all values for age 10-19 were represented as 14.5. However, this reduced the accuracy of both models to 65%, no better than a random guess. This may mean that the values are skewed inside the bins - it would be good to have the original data.

Naive Bayes

I ran 'OptimizeHyperparameters' on the Naive Bayes model, which returned a best performance with a kernel distribution, a kernel width of 0.60478, and a triangle. Since the only numeric variable is degree of malignancy, and the reported accuracy gain was less than 1%, I discounted this result and used 'mvmn' predictors for all variables, and made degree of malignancy a categorical variable, since the only values that appear are 1,2, and 3.

Random Forest

As well as using TreeBagger, I experimented with fitcensemble, due to the 'OptimizeHyperparameters' functionality and the option to bin numeric variables, but found that it greatly increased computation time to 8.0611s to train a basic model, and 368.9084s to optimise hyperparameters. It returned a maximum number of splits of 38 and a minimum leaf size of 12, however, I found that these did not significantly improve my model.

I also used the Classification Learner App on Matlab to obtain some baseline results.

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.