CITY UNIVERSITY OF LONDON
EST 1894

Comparison Between Random Forest and k-Nearest Neighbors For Breast Cancer Tumour Diagnosis

By Orestis Makris and Radi Khasawneh

Motivation and Description of the Data

•We have chosen a diagnostic breast cancer data set with numerical variables to seek to predict whether a growth is malignant or benign based on measurements taken from photographs of the growths.

•This is a therefore classification problem and we have selected examples of "lazy", locally weighted and eager, globally based classification models to provide an interesting comparison point.

•The nature of the dataset means that high levels of accuracy and low instances of false negatives will be the most desired result. We have decided to echo the approach taken by Saygili (2018) with the aim of increasing accuracy and F1 scores. We have also made reference to other papers which have taken a more wide ranging approach in the breadth of models used (including the two we selected).

Exploratory Data Analysis

•The dataset chosen was the Wisconsin breast cancer diagnostic sample published on the University of California Irvine machine learning database.

•Continuous variables have been extracted from fine needle aspirate images and classified into 30 variables (excluding the redundant numerical identifier column and categorical diagnosis column with was label encoded in Python). The data set is relatively small with 569 observations and we observed a slight mismatch in the target class, with a 62% to 37% split between malignant and benign tumors. This will have to be taken into account when splitting training and testing data.

•Additionally, initial column summaries and histograms revealed a natural correlation exists in the variables with a predominant right hand skew demonstrated in the column by column histograms which is evident in both classes of data.

•We also observed outliers that deviated hugely from the mean in three columns, which if left unchanged would negatively affect the performance of the K-NN algorithm.

•For all of those reasons, feature selection approach will be critical to maintaining a fair balance for performing the comparison. The high variance between the results of decision trees and the need for weak learners means that feature and hyperparameter selection will be important.

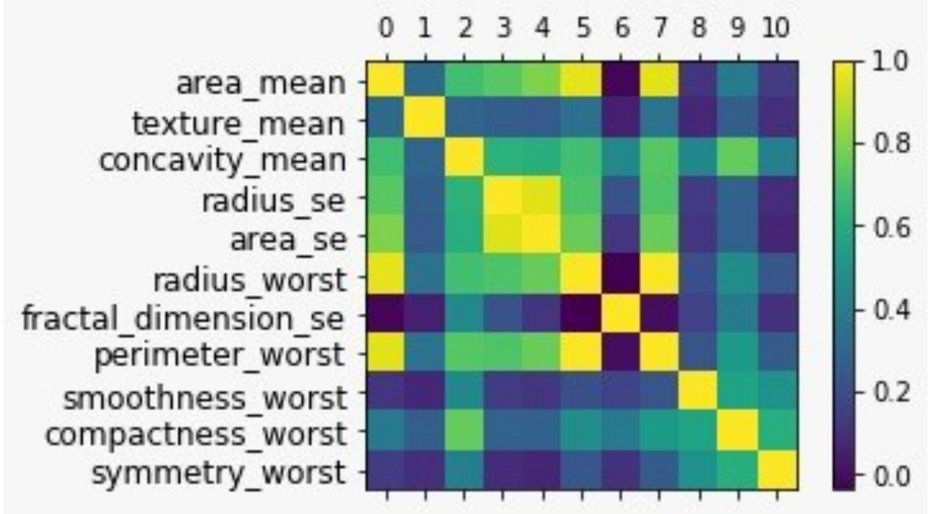
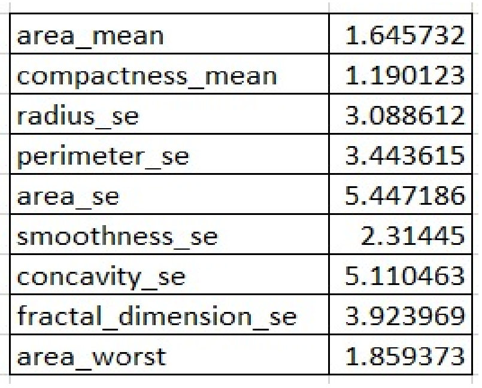


Figure 1: the correlation matrix for variables selected as best predictors



area_mean	1.645732
compactness_mean	1.190123
radius_se	3.088612
perimeter_se	3.443615
area_se	5.447186
smoothness_se	2.31445
concavity_se	5.110463
fractal_dimension_se	3.923969
area_worst	1.859373

Figure2: variable names and skewness columns with highest numbers chosen

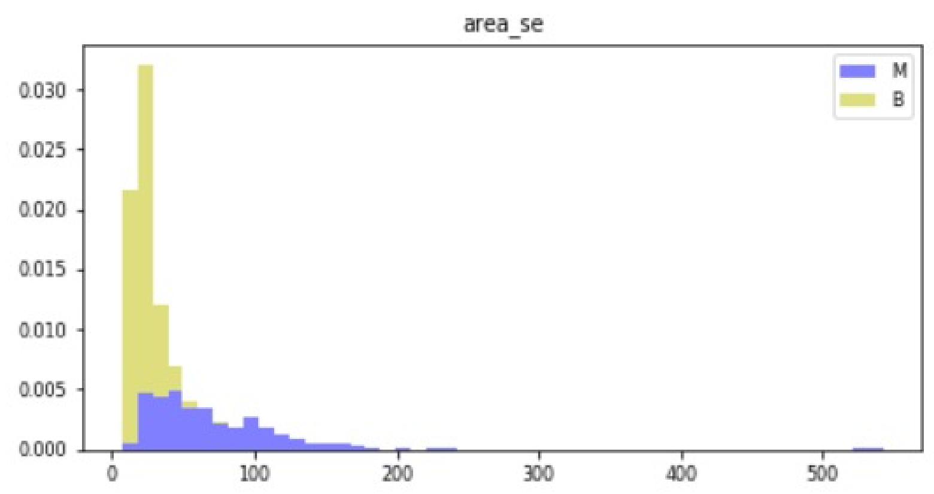


Figure 3: Histogram displaying most right skewed variable

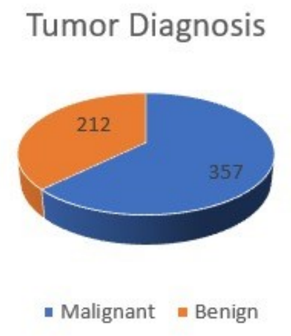


Figure 4: Tumour diagnosis shows imbalance in class labels

Model Comparison

Random Forest

•Decision trees have a long history with this dataset, in fact one of the first papers applying machine learning methods to it applied decision trees. The same author applied the C4.5 algorithm in order to deal with specific perceived problems in the result (Quinlan, 1995).

•Specifically, Quinlan cited the perceived weakness of the algorithm in dealing with and classifying continuous datasets and suggested an adapted approach to capture this difference.

•Bootstrap aggregation (bagging) through ensemble supervised decision tree ensemble analysis, can be expected to mitigate these effects and in fact the algorithm has met with high success and accuracy scores for this and similar datasets (Alam et al., 2019). This also negates the need to split the data in the training phase, although a high level of trees is recommended in order to ensure each feature is examined thoroughly.

Pros:

•Eager, global training, ability to counteract variance and need for early stopping. Easy to extract error and predictor estimates by default.

•Taking the average results will smooth the variance and to some extent mitigate the effect of decision trees, one of the main selling points of random forest.

Cons:

•Computationally expensive and (comparatively) slow, More difficult to interpret results and tune hyperparameters iteratively. Limited ability to manage hyperparameters (we have tried to mitigate this by variations in tuning).

•More difficult to interpret results and tune hyperparameters iteratively. Limited ability to manage hyperparameters (we have tried to mitigate this by variations in tuning).

K-NN

•By way of contrast, we have opted for an instance-based learning method that has achieved success on this dataset in the academic literature.

•This kind of "lazy" (Mitchell, 2017) learning method is categorised by delaying processing until classification, making it computationally efficient and locally agile (i.e. not generalizing the target function for the entire space but on an instance by instance basis).

•We have restricted ourselves to using distance measures rather than distance weights, because the normalisation of the data should reduce any negative effects of the "curse of dimensionality" (Bishop, 2009). This occurs because, unlike decision trees, the distance calculation occurs based on all points within the instance being modelled, rather than the relevant attributes for a splitting decision.

Pros:

•Simple and efficient, can adapt to new local events. Continuous and correlated variables may help with this dataset.

•In addition, the limited options make hyperparameter tuning very efficient (and well suited to Bayes optimisation).

Cons:

•Computation of all distances may stand against k-NN for this dataset, even where the dataset is normalised because of the likelihood of high importance features and many unimportant ones in many cases.

•Sensitive to unbalanced classifiers, noisy datasets and high dimensionality (leading to the need to take mitigating measures).

Hypothesis Statement and Methodology Description

Hypothesis:

•Although there is significant variance in the literature, we would expect random forest to perform better initially, but for k-NN to show the greatest improvement in the feature reduction run.

• We expect the two step hyperparameter tuning (Bayes optimisation and grid search) to show that a more exhaustive grid search will yield better results as iterations increase.

•Because the data can be characterised as high dimensional with low outputs there is a significant risk of the "curse of dimensionality" hitting k-NN, because it relies on a similarity metric and contrasts with random forest in that the distance between instances is calculated based on all attributes (Mitchell, 1997). We surmise that because all attributes are statistical measures of the same phenomenon, the most important predictors (or relevant attributes) will be drowned out while random forest benefits from weak learners in the cross validation process.

Methodology:

•As recommended by Saygili, we normalised the data and performed cv sampling to account for the imbalance in our classes. We have opted for a 70/30 percent training and test split and 10 fold cross validation, which is consistent with many approaches to this dataset (there is an exhaustive review of these in Yue et al., 2018).

•Rather than simply introducing feature selection in the pre-processing stage as is common, we utilised the approach detailed in Bhatia et al (2019), who ran two separate datasets through the fitting process to determine the effect on each algorithm with or without feature selection.

•We will compare out of the bag prediction measures (to approximate gain ratio based selection) and the minimum redundancy maximum relevance algorithm to determine the best feature reduced dataset.

•We have opted for a predictive and efficient Bayes optimisation for our hyperparameter tuning and will then compare the performance of our models based on both accuracy, F1 score and misclassified tumours to more accurately reflect the real world application of the data (i.e. the risk that the disease goes undiagnosed). Bayes optimisation uses the prior of an objective function and the uncertainty of the associated posterior (acquisition function) to predict the best hyperparameters, rather than a manual grid search that simply cycles through all the possibilities.

Experimental Results and Future Work

Experimental Results:

•We first compared feature predictors based on random forest and FSCMRMR method after which we chose 11 columns each as our most important predictors (above a certain threshold). After analysing the selections, the FSCMRMR were less correlated and therefore we opted to use those for our feature reduced run.

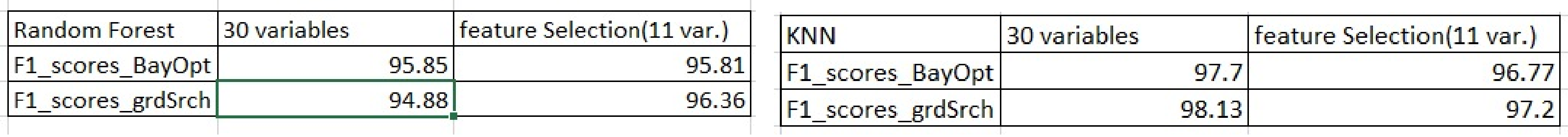
•In terms of Bayes optimisation with the full dataset, k-NN performed the best, with an F1 score of 97.7% (we have focussed on F1 scores because the balance of precision and recall more accurately reflects performance in the context of the domain). Random forest, at 95.85%, performed slightly worse. In the manual grid search, k-NN surprised us by returning 98.13%, against random forest with 94.88%. We cannot conclude that dimensionality harmed k-NNs relative performance as we had expected.

•With the feature reduced dataset, k-NN performed best with the grid search but they were roughly similar and outperformed random forest which again underperformed with Bayes optimisation. For random forest we recorded an F1 score of around 96 percent (see figure), which probably means that on balance we thought random forest underperformed our expectations while k-NN's results were near the high end of what we thought was achievable.

•In terms of hyperparameter tuning for k-NN, city block was chosen as the preferred distance twice (once in each dataset), with Euclidean being chosen along with 10 neighbors for the first Bayes optimisation run. The feature reduced dataset optimised at 16 neighbors under Bayes, with 3 being chosen for both grid searches.

•In terms of hyperparameter tuning for k-NN, city block was chosen as the preferred distance twice (once in each dataset), with Euclidean being chosen along with 10 neighbors for the first Bayes optimisation run. The feature reduced dataset optimised at 16 neighbors under Bayes, with 3 being chosen for both grid searches.

•For random forest there was a big deviation in feature selection and learning cycles, with leaf size staying relatively stable. In matimum features/predictors to sample we saw one selected twice for bayes optimisation (full/feature selected dataset) and a larger number for grid search (17 and 23 respectively).



Random Forest	30 variables	feature Selection(11 var.)
F1_scores_BayOpt	95.85	95.81
F1_scores_grdSrch	94.88	96.36

KNN	30 variables	feature Selection(11 var.)
F1_scores_BayOpt	97.7	96.77
F1_scores_grdSrch	98.13	97.2

Figure 5: F1 scores for KNN and Random Forest for full and feature selected dataset, with Bayes optimisation and grid search

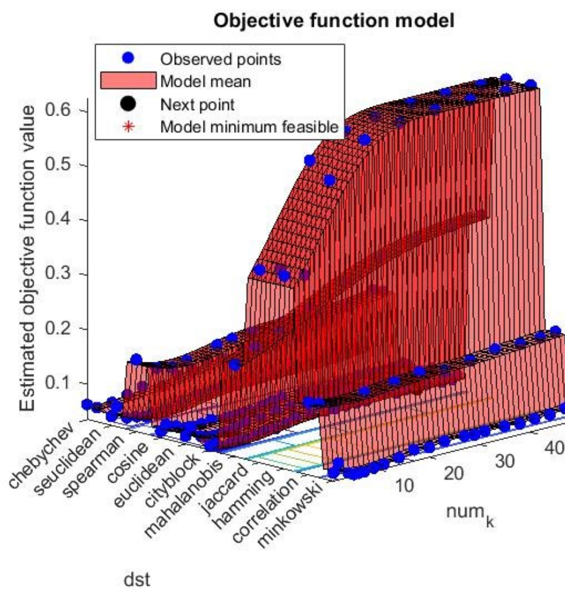


Figure 6: k-NN full dataset Bayes optimisation objective function formulation

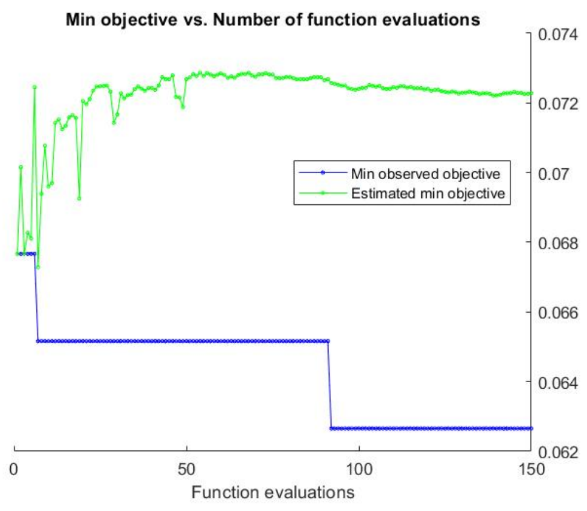
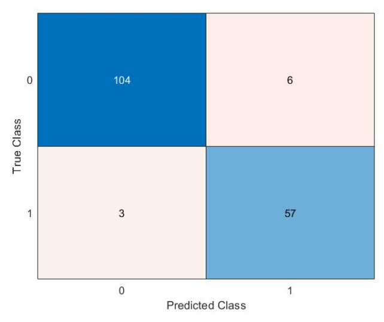


Figure 7: Random Forest feature selection minimum objective v function for Bayes optimisatin

Figure 8: Random Forest Confusion Matrix, full dataset, Bayesian optimisation result



Future Work:

•Our main takeaway from this project was that while this dataset achieves a high level of accuracy at the start point, a highly structured and technical approach is required to tangibly improve the model.

•More sophisticated feature selection, such as recursive feature elimination (RFE) that ranks and then eliminates features either before or after sampling would be instructive, as well as simply selecting a small amount of features (automated feature selection suggested 3 and 4 variables for k-NN and random forest respectively).

•We would consider experimenting with subspace sampling and selecting a smaller amount of weighted features to reduce the computational load, as used by Alam et al (2019) for feature selection. In general, a much deeper and more refined cross validation technique would also help us more closely replicate the academic results.

•An ensemble of k-NNs was also considered but the literature was patchy and inconclusive on whether this would lead to any improvement (Zhang, 2019).

•We would actually be interested in trying a support vector (SVM) because we have two classes that would lend themselves to this hyperplane identification approach (Yang et al, 2015), which could even be expanded into a hybrid algorithm with an ensemble of decision trees, as applied by Sivakami (2015).

•Lastly, we would consider applying a cost matrix to increase weight on misclassified tumours to more closely reflect the nature of the domain and practical application of the dataset.

References

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Alam, M., Rahman, M. and Rahman, M. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 15, p.100180.

Bhatia, S., Tiwari, S., Mishra, K. and Trivedi, M. (n.d., 2019). *Advances in Computer Communication and Computational Sciences*.

Saygili, A. (2018) Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers, *ISVOS Journal* 2018, 2(2):48-56,

Elgedawy, M. (2017). Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes. *International Journal Of Engineering And Computer Science*.

Mitchell, T. (2017). *Machine learning*. New York: McGraw Hill.

Sivakami, K. (2015). Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model, *International Journal of Scientific Engineering and Applied Science*, 1(5)

Managasarian, O. (1998). Classification and feature selection applied to breast cancer diagnosis. *ACM SIGBIO Newsletter*, 18(3), pp.8-8.

Quinlan, J. (1996). Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, 4, pp.77-90.

Yan, K. and Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, 212, pp.353-363.

You, W., Yang, Z. and Ji, G. (2014). Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination. *Expert Systems with Applications*, 41(4), pp.1463-1475.

Yue, W., Wang, Z., Chen, H., Payne, A. and Liu, X. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs*, 2(2), p.13.

Zhang, Y., Cao, G., Wang, B. and Li, X. (2019). A novel ensemble method for k-nearest neighbor. *Pattern Recognition*, 85, pp.13-25.