# Comparison of Machine Learning Methods for Breast Cancer Detection

Introduction:

With breast cancer being one of the most prevalent cancers affecting women in the world today, any effort to detect it represents a public health investment with an potentially enormous payoff. As with any cancer, early detection is key, and therefore, the focus of any such effort must be on the earliest stages possible. As computation power increases exponentially and new statistical methods are devised, we are presented with an ever-expanding machine learning tool kit which can and must be leveraged to these types of problems. While the scope of this analysis is rather limited, it is intended to serve as a demonstration of the possibilities of machine learning, not just in the medical field, but in society as a whole.

Methods:

Data were obtained from the Wisconsin Breast Cancer Database, with the data set containing 569 observations. Each observation is a sample of cells from a breast mass, with a dichotomous response – the mass is either benign or malignant. For each mass, 10 aspects of the cells sampled are measured for each mass, with the mean, standard error, and extreme value being calculated for each aspect of each mass. The result is a total of 30 values, which are used as the predictors for each classification method.

|  |
| --- |
| Radius |
| Texture |
| Perimeter |
| Area |
| Smoothness |
| Compactness |
| Concavity |
| Concave Points |
| Symmetry |
| Fractal Dimension |

Table 1: The ten cellular aspects that were used as predictors. For each of these aspects the mean, standard error, and extreme value are calculated for the sample of cells from each mass.

Univariate histograms were produced for each of the predictors, and then each predictor was examined again when split by the response.

Next, the data set was split into a testing set and training set, with 80% of the observations going to the training set, and 20% going to the testing set. This ratio was chosen in order to maximize the number of observations used to train each model, as this data set has a relatively low number of observations. A model was fit on the training data set for each of the methods that were used, and then the accuracy of the model was evaluated by using it to make predictions on the testing data and then comparing it to the actual testing data response. Finally, an ensemble method was employed in which predictions from each method were employed in a majority rule vote to determine the final ensemble prediction.

|  |
| --- |
| Generalized Linear Model |
| Linear Discriminant Analysis |
| Quadratic Discriminant Analysis |
| K-nearest Neighbors |
| Naive Bayes |
| Random Forest |
| Support Vector Machine |
| Ensemble |

Table 2: List of Supervised Methods used to classify data

Results:

Univariate:

THERE IS AN IAMGE HERE IN THE FINAL PAGE

Figure 3: Univariate Histograms of Each Predictor

Bivariate:

THERE IS AN IAMGE HERE IN THE FINAL PAGE

Figure 4: Bivariate Histograms of each predictor. Malignant in red, benign is superimposed in blue

Multivariate:

|  |  |
| --- | --- |
| Method | Accuracy |
| GLM | 98.25% |
| LDA | 99.12% |
| QDA | 99.12% |
| KNN | 98.25% |
| NB | 92.11% |
| RF | 100.00% |
| SVM | 98.25% |
| Ensemble | 100.00% |

Table 5: Classification Accuracies for each method

Discussion:

From Figure 3, it is apparent that the majority of values for each predictor are very close to 0, with a tail of much higher values. The higher values for each predictor are sparse and spread out, with the distribution of each displaying a large degree of skewness. Looking at Figure 4, it can be seen that values for each predictor are strongly separated into one class of the response or the other when the values are much greater than 0. While this does not seem to be the case for the lower values of each predictor, if the predictors are not strongly correlated, that is if each observation has at least one predictor with a high value, this data should be conducive to accurate classification.

Table 5 confirms this suspicion – all of the classification methods achieved at least 90% accuracy, with two – random forest and ensemble, achieving 100% accuracy! These results implicate the random forest method as the best classification method, however with the scope of this analysis being limited, this is a tenuous assumption to make at best. A fairer takeaway would be the ensemble method – as it incorporates all methods into its predictions (though it could be hindered by having several inaccurate methods used).

Limitations and Future Work:

Due to the relatively low number of observations in this data set, observations were given to the training set with higher priority than the testing set. This was necessary in order to ensure accurate models were fit, but as a result, the accuracy had a high degree of granularity as evaluated by the low number of observations in the testing set. Had there been significantly more observations in the testing set, it is unlikely that any method would’ve had exactly 100.00% accuracy.

Future work with this data set can involve unsupervised clustering methods, to compare against the supervised methods used in this analysis, as well as resampling methods to try and alleviate the granularity due to the low number of observations in the testing data set.

Nonetheless this analysis demonstrates the utility of machine learning methods applied in a medical context. It is not difficult to imagine a myriad of other applications for these methodologies.

References:

1. O. L. Mangasarian and W. H. Wolberg: “Cancer diagnosis via linear  
   programming”, SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
2. William H. Wolberg and O.L. Mangasarian: “Multisurface method of  
   pattern separation for medical diagnosis applied to breast cytology”,  
   Proceedings of the National Academy of Sciences, U.S.A., Volume 87,  
   December 1990, pp 9193-9196.
3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: “Pattern recognition  
   via linear programming: Theory and application to medical diagnosis”,  
   in: “Large-scale numerical optimization”, Thomas F. Coleman and Yuying  
   Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
4. K. P. Bennett & O. L. Mangasarian: “Robust linear programming  
   discrimination of two linearly inseparable sets”, Optimization Methods  
   and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).