Machine Learning Classification of Benin and Malignant Breast Cancer Tumors Using FNA Digitized Images

Amir Ghaderi – 500794236
Mezbah Uddin – 500793378

DS8004 – Data Mining

Ryerson University

Introduction

Tumors are an abnormal growth of tissue and serve no purpose to the human body. These tumors can be classified as either benign or malignant. Benign tumors are non-cancerous and do not invade neighboring cells. Benign tumors can be caused by a number of reasons: environmental toxins, genetics, diet, stress, local trauma, or infection. Benign tumors are non-threating to humans and often do not require surgical intervention. Benign tumors are usually placed on a watch list and monitored closely. Malignant tumors however are cancerous and do invade neighboring cells. These tumors are considered extremely dangers and if left untreated are considered to be fatal. Malignant tumors have the ability to multiply uncontrollably and spread to surrounding organs and tissue.

There are several distinct differences in appearance between benign and malignant tumors. For example, Malignant tumors have larger nuclei then benign tumors. Currently the industry standard method for classifying breast tumors (mass) is through a procedure called Fine-needle aspiration (FNA). Fine-needle aspiration is diagnostic procedure that is used to investigate tissue masses or lumps. It requires an extraction of part of the tumor through the user of a needle. The extracted tissue is then examined under a microscope and the tumor is classified as either benign or malignant. However, 25% of the time this procedure yields an inconclusive result. In these cases (a quarter of the time) the tumor needs to be removed using an invasive surgical procedure.

Researchers at the University of Wisconsin took digitalized images of breast mass while performing fine needle aspiration and studied the inconclusive classification cases. The researchers used 56 of the FNA digitizable images and attempted to classify the tumors using

machine learning algorithms. The researchers obtained the actual labels of the tumors using surgical procedures and after classification obtained an accuracy of 75%. The university of Wisconsin then published a dataset contained FNA digitalized images from 569 inconclusive FNA diagnostics.

Our goal of this project is to work with these FNA digitalized images and use machine learning classification algorithms to classify these tumors as either benign or malignant. We hope to be able to obtain a classification accuracy that is higher than the initial study which obtained an accuracy of 75%.

Data Source

The dataset comes from the UCI Machine Learning Repository. The dataset it titled "Breast Cancer Wisconsin Diagnostic dataset and can be obtained from the following link: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

The dataset contains 569 records, 32 attributes, and 10 features. The features are as follows:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)

8. concave points (number of concave portions of the contour)

9. symmetry

10.fractal dimension ("coastline approximation" - 1)

There are two classes in this dataset, Benign and Malignant. The class distribution is slightly skewed contained 67% benign and 33% Malignant. There are no missing attributes of values.

Data Preparation and Dimensionality Reduction

The dataset was imported into R as a data frame from a comma separated values file. The dataset originally did not contain column names, therefore we appended the original names of the features from the appendix of the UCI data source. Once the dataset was imported and labeled we normalized the dataset using the max/min criteria (range 0 to 1). After normalization, we explored the associations among our attributed. We were able to accomplish this by creating a correlation matrix, which revealed that there exists a strong linear association between the attributes. Therefore, dimensionality algorithms are required in order to minimize the “curse of dimensionality”. Since our variables contained strong linear associations it is optimal to use a linear dimensionality reduce algorithm for example principle component analysis. We decided to use 10 principle components which provided us with a cumulative variance proportion of 95.15%. Then we appended the class labels to the PCA transformed dataset with 10 principle components. Finally, we converted our class feature into a factor in order to pass the dataset through various machine learning algorithms.

Classification

We decided to pass our dataset through 5 machine learning algorithms and record the accuracy of each classifier. We decided to use five classifiers in order to achieve a distribution of accuracies. Therefore, we would be able to compare and contrast the accuracies among various classifiers. The first algorithm that we used is support vector machines (SVM) with linear, polynomial, radial, and sigmoid kernels. The second algorithm that we used was K nearest neighbor (knn) with k values of 1, 3, 5, 11, 19, 21. The third algorithm that we used was a random forest decision tree with tree values of 100, 200, 300, and 400. The forth algorithm that we used was a decision tree (bagging). The final algorithm that we used was multilayer perceptron with 1 hidden layer. For the random forest and SVM algorithms we applied 5 fold cross validation in order to see the generalized performance of the classifiers.

SVM

Support vector machines is a supervised machine learning algorithm that attempts to separate the classes by maximizes the distance between classes. The data points that are on the fridge of each cluster are considered to be the support vector machines. These support vector machines are the bases of this classifier.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Linear	95.69%	97.54%	97.24%	97.32%	97.54%
Polynomial	94.18%	95.29%	95.89%	95.56%	95.79%
Radial	95.66%	96.43%	95.78%	96.34%	96.84%
Sigmoid	91.9%	94.12%	93%	90.81%	90.34%

As you can see from the table above the linear kernel on average out performs the other kernels. This is most likely the case because our classes are linearly separable using a higher dimensional hyperplane. In addition, the sigmoid kernel performed the worst on our dataset because it used a hyperbolic tangent function that is nonlinear.

KNN

K nearest neighbor is a supervised machine learning algorithm that classifies unlabeled data points based on the points k nearest neighbors. The most common distance function that is used is Euclidean distance.

# of Nearest Neighbor	**Accuracy**
K=1	92.98%
K=3	93.85%
K=5	95.61%
K=11	94.69%
K=19	98.24%
K=21	98.24%

As you can see from the table above when k=21 the algorithm yields the highest accuracy. This is most likely the case because the optimal number of neighbor is usually equal to the square root of the number of observations in your dataset.

Random Forest

Random Forest is a supervised machine learning algorithm that classifies the unlabeled data points by passing them through a multitude of decision tress. This ensemble algorithm outputs the mode or mean prediction of individual trees.

# of Trees	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
100	96.47%	94.6%	93.77%	94.86%	94.91%
200	97.15%	97.5%	96.67%	95.63%	95.08%
300	95.46%	94.27%	94.7%	94.74%	94.38%
400	93.97%	95.24%	95.84%	95.40%	95.26%

As you can see from the table above the optimal result is achieved when the number of trees is equal to 200. This is the case because when we increase the number of trees passed 200 the model tends to overfit and therefore generalization error increases.

Bagging

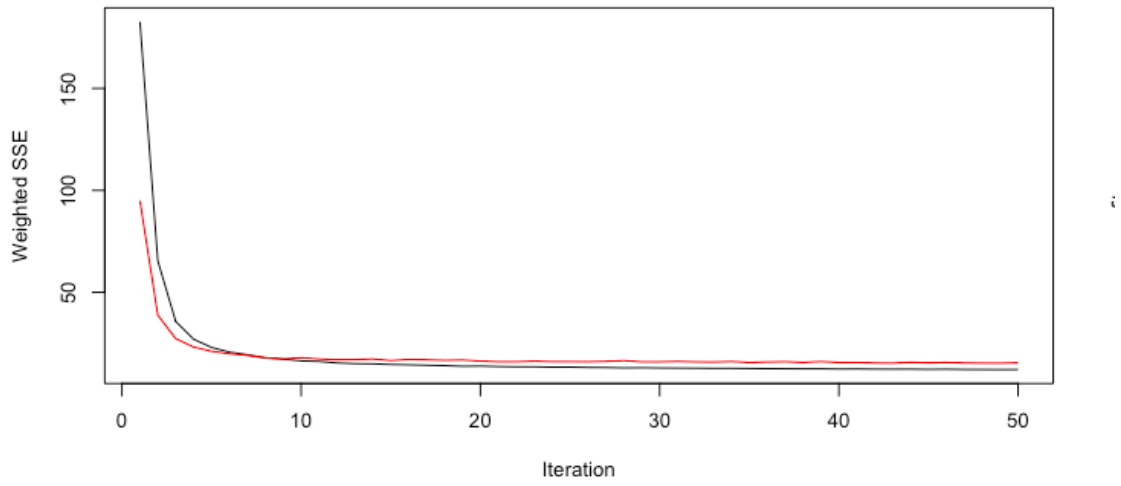
Bagging is a supervised machine learning algorithm that classifies the unlabeled data points by taking subsets of your dataset and passing it through classifiers. This ensemble algorithm outputs the mean prediction accuracy among all the subsets of your dataset.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
98.84%	95.56%	94.84%	95.05%	95.44%

As you can see from the above table the bagging algorithm achieved an average percent accuracy of 95.80%. This result was achieved across 5-fold cross validation iterations.

Multi-Layer Perceptron

Multi-Layer perceptron is a supervised machine learning algorithm that classifies the unlabeled data points by mapping them to a class. This algorithm using a feedforward artificial neural network that updates weights in each epoch.



As you can see from the figure above, as you increase the number of iterations the classification error convergences to 98% accuracy.

Conclusion

In conclusion, we set out to classify breast cancer tumors using FNA digitalized images and achieve a classification accuracy greater than 75%. Through the implementation of various machine learning classification algorithms, we achieved the follow maximum accuracies: {SVM: 97.54%, KNN: 98.24%, Random Forest: 97.50%, Bagging: 98.84%, MLP: 98.00%}. As you can see the algorithm that classified out dataset with the highest accuracy was bagging.

Since this can potential be used as a medical diagnostic tool, it is paramount that the error rate is extremely low. Since we are dealing with patients' lives, a false negative case can potentially be fatal.

References

[1] https://dollar.biz.uiowa.edu/~nstreet/research/cc97_02.pdf

[2] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[3] https://en.wikipedia.org/wiki/Fine-needle_aspiration

[4] https://www.r-project.org/

Appendix: