Breast Cancer Classification project

Justin Chen

March 2022

1 Description of Data

An analysis of the "Breast Cancer Wisconsin (Diagnostic) Database" a joint study created by the University of Wisconsin's Computer Science Department and General Surgery Department(1). The actual data set was downloaded from "UCI Machine Learning Repository"(2). The goal of the study was to measure the properties of 569 different breast cells nuclei. Using digitized imaging on a "fine needle aspirate" a technique that uses a needle to collect samples from a bump or lump on the breast. The list of these measurements are as follows:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area in radius lengths
- compactness (perimeter^2)
- smoothness (local variation / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points
- fractal dimension (concave portions of the contour)
- symmetryastline approximation" - 1)

The study also noted whether the cell was "Malignant" or "Benign". In other words, whether the patient had breast cancer or not respectively. The actual data set had 32 variables, however the analysis will only observe 10, 1 being categorical (diagnosis) and the rest being real valued. This is because of the data description given by the authors, only ten variables out of the data set are clearly classified. Using "Kaggle"(3) the appropriate variables are named in the data set.

2 Problem

To classify the cell as malignant or benign using the real valued breast cell nuclei measurements. This will require an analysis on how exactly the two classifications differ in terms of cell nuclei properties. Note that the goal is prediction as oppose to modelling

3 Techniques

Since all the variables are real valued and the goal is classification, multiple techniques are available. The first technique used will be "K nearest neighbours"(4). This machine learning technique uses a training set and a test set, made up of observations from the data set. The training set will have 80 percent of the observations, made up of labelled observations that will "teach" the k-NN algorithm how to label unlabeled observations. The k-NN algorithm will then be used on the test set which is the remaining 20 percent of the data set made up of unlabeled observations which the algorithm labels. The neighbourhood is a distance measure of similarity between observations. The k value determines how close observations must be to be classified in the same group. The algorithm will be run with multiple "k" values. The smallest k-value with optimal classification results will be used.

The k-NN algorithm results will be compared to the multiple variations of the clustering technique. Clustering in general is putting classifications on a data set that is treated as unlabelled. The first type of clustering used will be agglomerative hierarchical clustering(5). This method takes every observation and separates them into clusters, we then take the two nearest observations and form a single cluster. This step is reiterated until all observations of close proximity are in one cluster. The Euclidean distance will be

used to measure to measure the similarity/distance between two observations: $d(x_i, y_j)=\sqrt{\sum_{m=1}^{M}(x_{im} - x_{jm})^2}$(6) There are multiple ways to measure the distance between clusters, the ones of interest will be: single, ward.D and ward.D2. Single takes the minimum distance between two points in a cluster, ward.D uses a weighted sum of distances and ward.D2 uses the square root of the weighted sum of distances.(5). The ward.D and D2 methods are usually the most accurate as they reduce in cluster variances. The dendrograms will be used to evaluate how useful each method is. An ideal dendrogram will have proper spacing of individual entities without so much overlap (chaining). A classification table will then be outputted using the optimal linkage method with a branch cut of two to represent the two classes of interest (benign and malignant).

The final method used will be "K-means and K-Medoids clustering"(7). Both select k clusters that minimize the distance between the centre of the cluster and the other points within the cluster. The cluster centres being means or medoids. The K-means algorithms need specific k values chosen before hand, this choice will be selected using the silhouette plot. k=2,3,4 will be tested, whichever plot has a combination of high average length, minimal observations in the wrong cluster, and overall which k gives us the best information. k=2 is a valid hypothesis since the data has two classes of interest so having two clusters would be perfect for the data. The k-means and k-medoids algorithm will be run for the optimal value of k along with a classification table.

4 Results

K-Nearest Neighbours produced multiple results depending on the k value. In particular the best classification results are produced with k=1,k=2 and k=3. For k=1:

	B	M
B	234	120
M	5	176

For k=2:

	B	M
B	220	134
M	6	175

For k=3

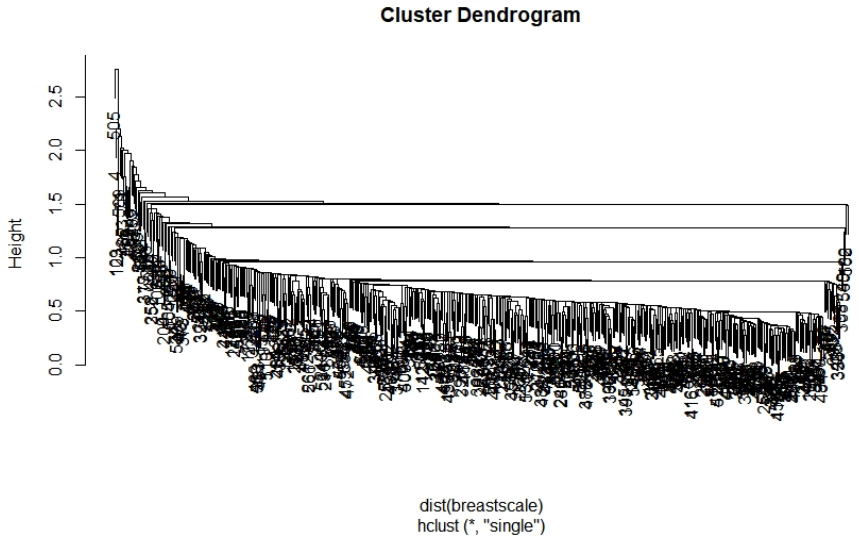
	B	M
B	241	113
M	3	178

For k=1 there are 125 miss-classifications, k=2 there are 140 miss-classifications and for k=3 there are 116 miss-classifications. While k=3 has the least miss-classifications its only by a margin of 9 observations. The general ruling is if the k values yield similar results then the smallest k is taken, so in this case k=1 is best choice for KNN. However, even with k=1 being the optimal choice the classification rate is still concerning. In total there are 125 miss-classifications. In addition the $k > 3$ algorithms are not impressive in terms of correctly classifying cells.

As k gets larger the miss-classification rate only increases, until it reaches 354 miss-classifications at k=7.

	B	M
B	0	354
M	0	180

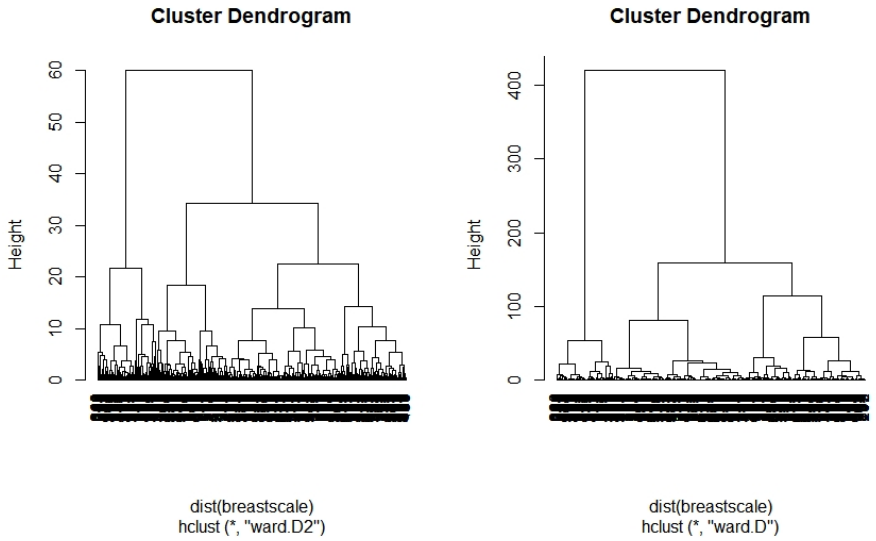
In summation, the K-NN algorithm wasn't optimal, other methods will prove to be more effective. Agglomerative Hierarchical Clustering will be the next method. Ward.D2 turned out to be the optimal linkage method as demonstrated by the corresponding dendrograms. For "Single" linkage:



This is not an optimal Dendrogram as there is chaining amongst observations. The single linkage method is not ideal and is actually the worst linkage

method.

Ward.D2 will be compared with the "Ward.D" method using hang=-1 resulting in:

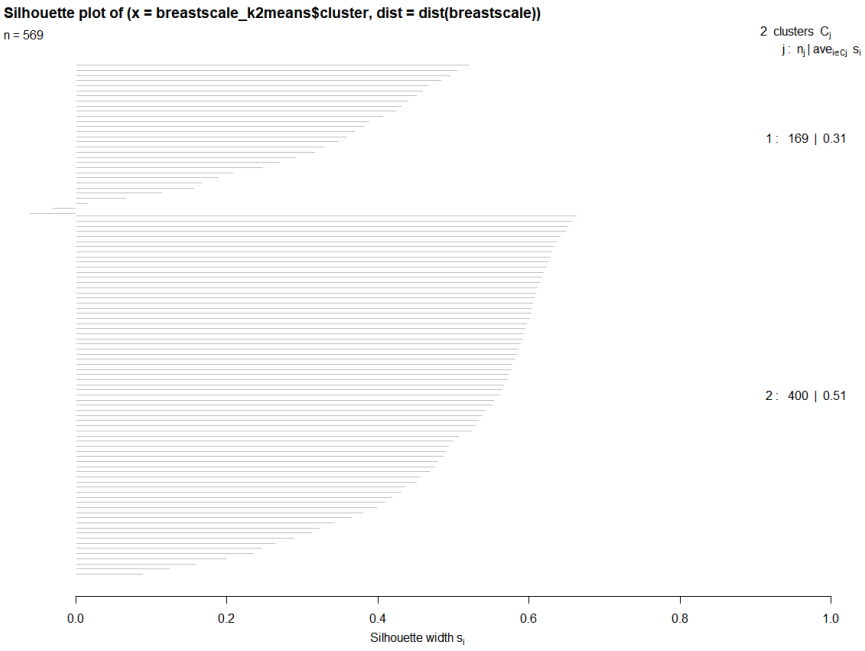


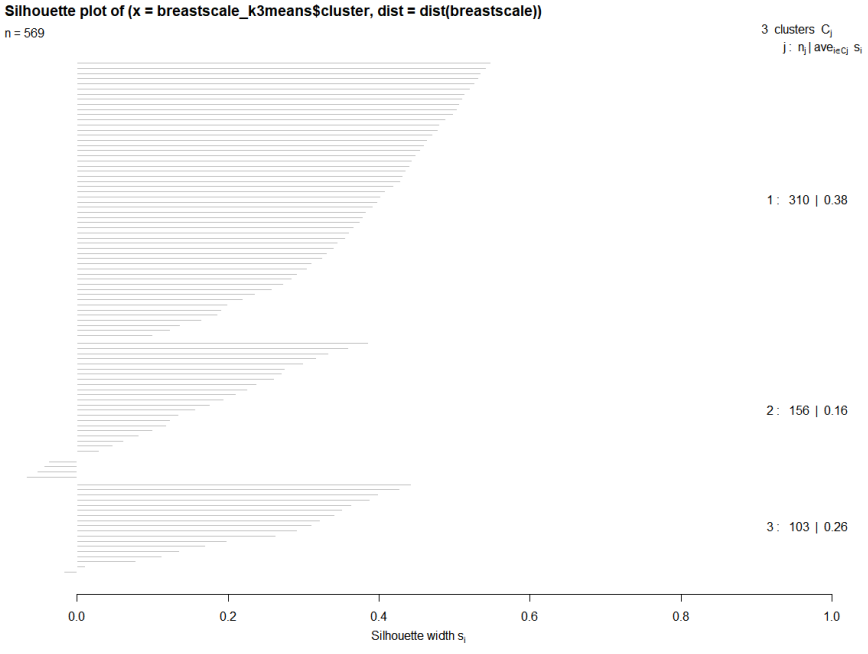
Ward.D2 in general is a more refined version of Ward.D. So Ward.D2 will be will be the optimal linkage method. The classification table for Ward.D2 is

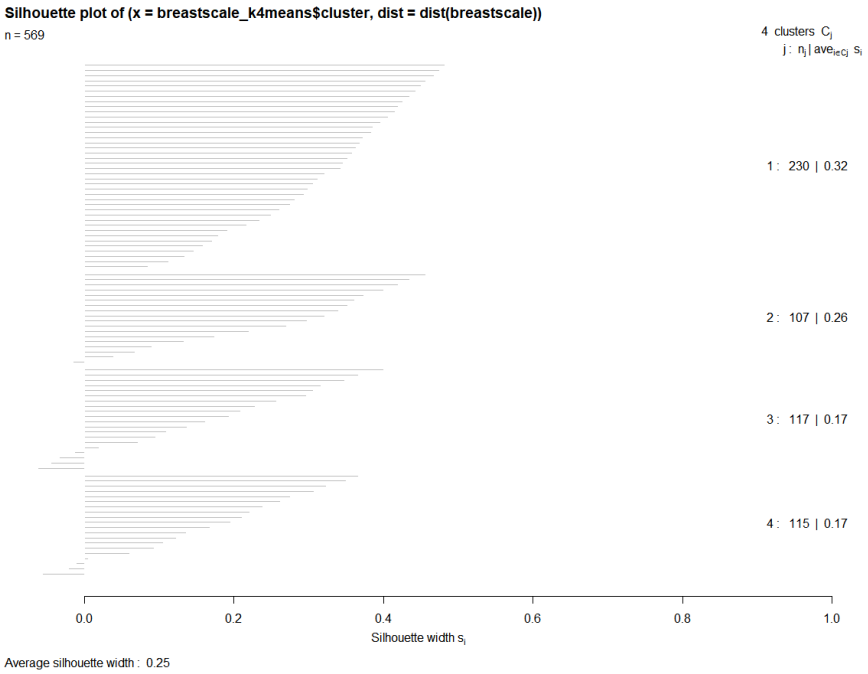
	M	B
B	0	357
M	109	103

There are in total 103 missclassifications. A slightly better result then the K-NN algorithm however this change isn't extreme. Both methods are comparable in terms of effectiveness.

K-Means and K-Medoids both proved to be more effective compared to the previous two. First compare the silhouette plots for k=2,3, and 4


Silhouette plot of (x = breastscale_k2means$cluster, dist = dist(breastscale))
n = 569
2 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 169 | 0.31
2 : 400 | 0.51
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.45


Silhouette plot of (x = breastscale_k3means$cluster, dist = dist(breastscale))
n = 569
3 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 310 | 0.38
2 : 156 | 0.16
3 : 103 | 0.26
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.3



k=2 gives the best plot. The average width of 0.45 significantly shorter then k=3 and 4 indicating more accurate observation allocation to the clusters. Two clusters makes sense given the two classes of interest.

The classification for k=2 is as follows:

	M	B
B	4	353
M	165	47

A noticeable difference in miss-classification rate. Here we only have 51 miss-classifications which is almost half the amount of the other two methods. The K-Medoids results were very similar to K-means, k=2 being the optimal value and the classification table being

	M	B
B	5	352
M	172	40

These results are slightly better having only 45 miss-classifications. However its clear that both of these methods are much more effective then K-NN and Hierarchical clustering.

5 Conclusion

Out of the three methods discussed the optimal method is K-means/K-medoids. Due to these methods having the best classification rate. Proving to have half the amount of miss-classifications as the other results and in general classifying about 90 percent of the observations correctly. While this is effective, given more time and capacity other methods such as the neural network are of interest.

6 Bibliography

text(1) Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, WI 53792 wolberg '@' eagle.surgery.wisc.edu

(1) W. Nick Street, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 street '@' cs.wisc.edu 608-262-6619

(1) Olvi L. Mangasarian, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 olvi '@' cs.wisc.edu

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

(3)https://www.kaggle.com/uciml/breast-cancer-wisconsin-data?select=data.csv

(4)Sharon McNicholas lecture 7 notes

(5)Sharon McNicholas lecture 8 notes

(6)Applied Multivariate Statistical Analysis (6th Edition)by Johnson and Wichern, page 30.

(7)Sharon McNicholas lecture 9 notes

(8) Sharon McNicholas lecture 10 notes

(9) Gaussian parsimonious clustering models. Pattern Recognit. 28,Celeux, Gilles Govaert, Gerard (1993).