Classification of the Breast Tissue using Linear Discriminant Analysis

A PROJECT REPORT

submitted by

Srikanth Gadicherla (546195)

for the course

MS-E2112 – Multivariate Statistical Analysis

Master of Science in **Machine Learning** and **Data Mining**

School of Science
AALTO UNIVERSITY

6th APRIL 2016

1. RESEARCH QUESTION

In this project, the breast tissue dataset was taken from the UCI repository [1]. The task at hand was classification of the given samples into their corresponding categories. The linear discriminant analysis was the technique used for accomplishing the task.

2. DATA INSPECTION:

This multivariate dataset contained 106instances of breast tissue with the ten attributes namely I0 Impedivity at zero frequency, PA500 phase angle, HFS high-frequency slope, DA impedance, Area under spectrum, A/DA normalized area, MAX IP, DR distance, P length of spectral curve. The impedance (~resistance) of the freshly excised breast tissue were measured at different frequencies of 15.625, 31.25, 62.5, 125, 250, 500, 1000 KHz. The data samples were used to classify them into six groups i.e., carcinoma, fibro-adenoma, mastopathy, glandular, connective and adipose.

2.1 Univariate Data Visualization:

The data was visualized with the univariate component-wise scatter plots as the data was real continuous data. Figure 1 shows the scatter plot with the different classes with distinct colors with the combination of data distributions on the principal diagonal of the plot. The visualization was accomplished using the R's ggplot2 package.

Initial thoughts from the plot is that because of the less data points the univariate distribution doesn't reveal any distribution which could be assumed for the next steps of the analysis. Next, most of the groups in the data can't be separated from the plots except the groups of adipose (salmon color) and connective (green color) and for some plots even the carcinoma (yellow color) group can separated while the other classes are seen over-lapping in most of the plots. The class adipose can be seen to be having outlier data samples.

The attributes I0-P and DA-DR are correlated respectively. This can be confirmed from the correlation plot which is plotted next.

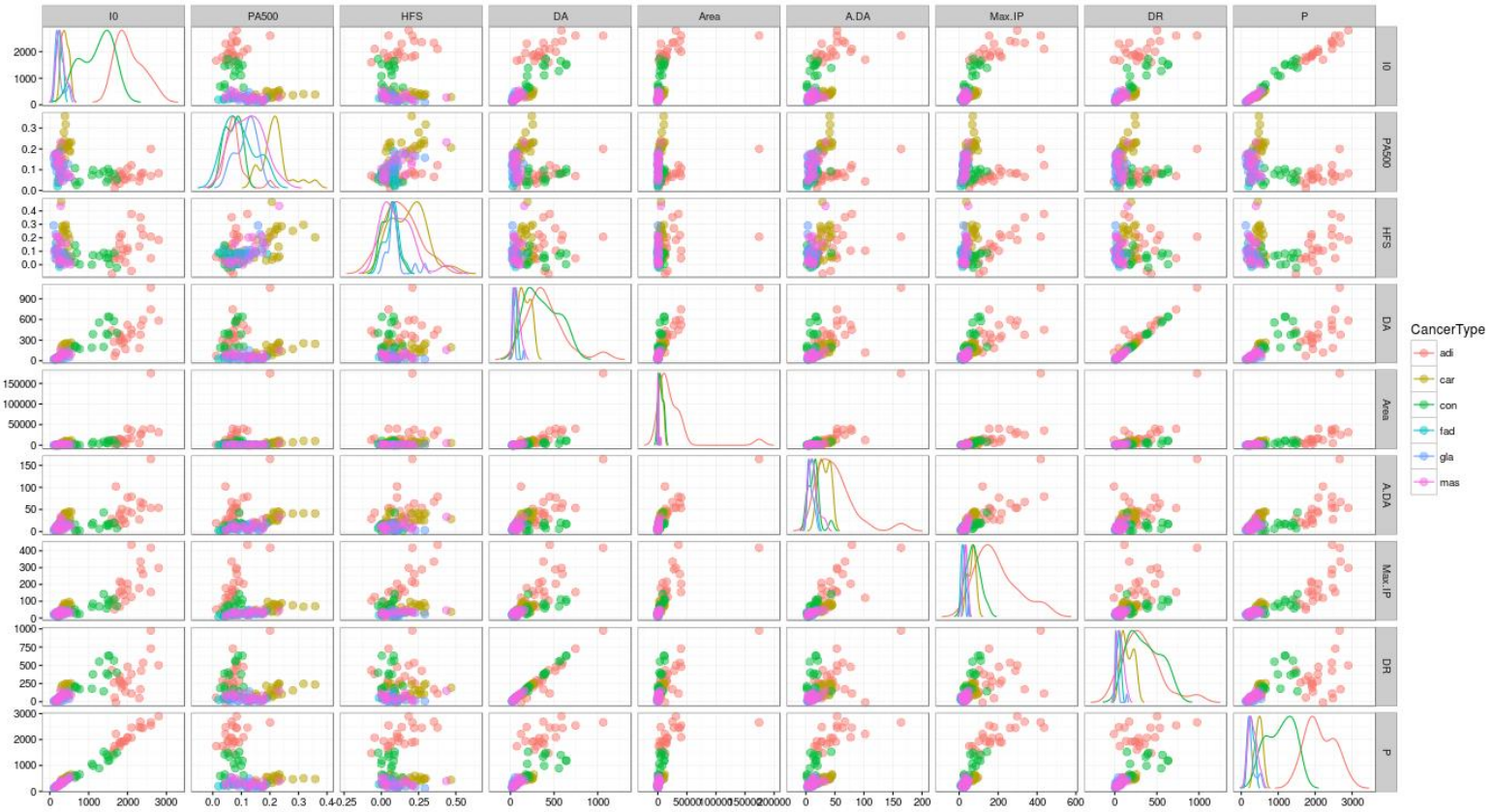

Figure 1: Scatter plot of the data.

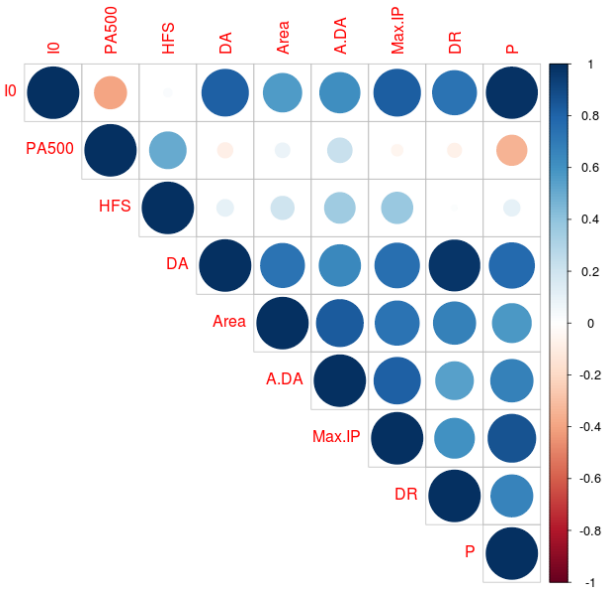

Figure 2: Bubble plot of the correlation matrix.

Next, the bubble type plot correlation plot is presented for the data. As told above, it can be seen that attributes I0-P and DA-DR are highly correlated with correlation coefficients 0.988 and 0.974 respectively.

The next important aspect is to plot the histograms of the continuous attributes of the data.

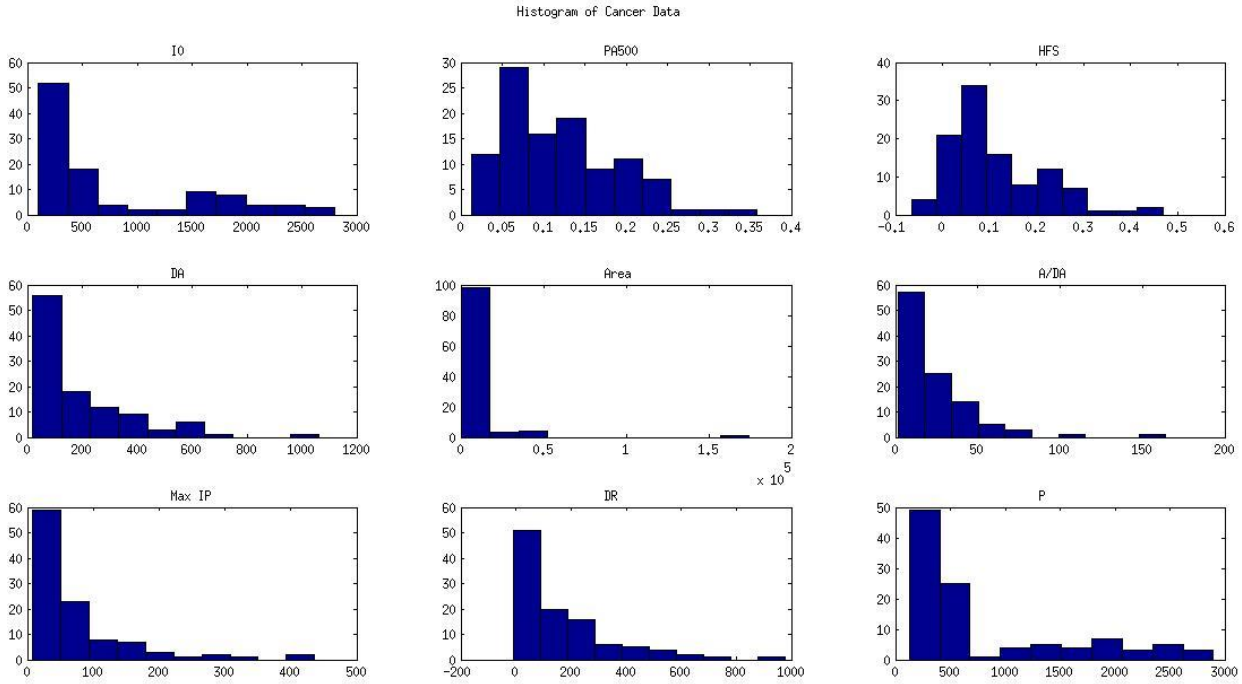

Figure 3: Histogram of the data.

The observation from as told from the scatter plot is that most of the attributes are skewed having no proper distribution. For the attributes Area, A/DA, DR, Max IP and DA have the outlier points.

It can also be seen from the histogram that range of the attributes is highly varying from 10^{-1} to 10^{+3}. So it's logical to present the box plots for all the attributes of the data.

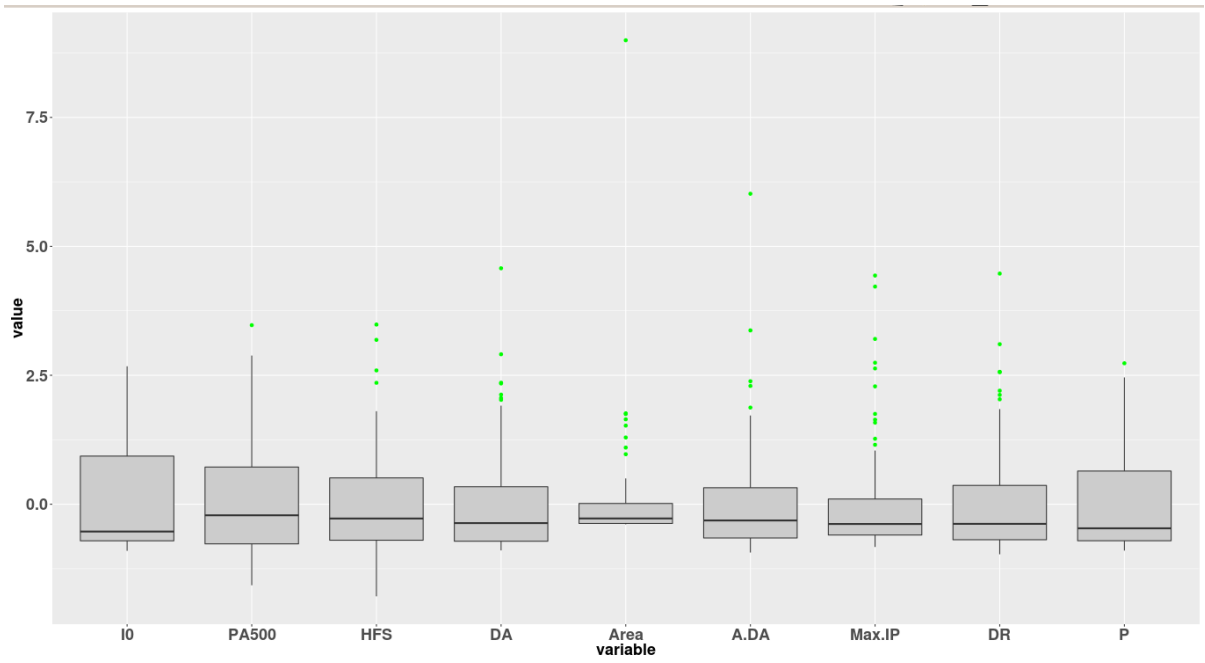

Figure 4: Box plot of the data.

Evident from the histogram that the long tails for the attributes in the histogram, the box plots has many outlying points. As the data had varying range, the data was normalized to get a sensible box plot.

Last, the data statistics were found out. They included mean, maximum, minimum, median and variances of all the attributes.

Statistic/ Attribute	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
Mean	784.25	0.12	0.11	190.57	7335.16	23.47	75.38	166.71	810.64
Max	2800.00	0.36	0.47	1063.44	174480.48	164.07	436.10	977.55	2896.58
Min	103.00	0.01	-0.07	19.65	70.43	1.60	7.97	-9.26	124.98
Median	384.94	0.11	0.09	120.78	2219.58	16.13	44.22	97.83	454.11
Variance	568440.72	0.00	0.01	36405.19	345228076.26	545.44	6617.15	32873.16	582198.20

Figure 5: Statistics of the data.

2. METHODS IMPLEMENTED and RESULTS

Linear discriminant analysis was the method applied for classifying the breast tissue for cancer cells. It deals with finding the direction in the data space where the group separation is maximum and at the same time intra-group dispersion is minimum.

The result for the data after dividing the dataset into 80% training and 20% test data is 75% accuracy.

		Actual Output					
		adi	car	con	fad	gla	mas
	adi	7	0	0	0	0	0
	car	0	0	0	0	0	0
Predicted Output	con	0	0	3	0	0	0
	fad	0	0	0	2	0	1
	gla	0	0	0	0	3	2
	mas	0	1	1	0	0	0

Figure 6: Prediction on test data shown as confusion matrix.

The cross-validation gives the accuracy of 66.9% on the whole dataset and is shown below.

		Actual Output					
		adi	car	con	fad	gla	mas
	adi	20	0	1	0	0	0
	car	0	18	0	1	0	2
Predicted Output	con	2	0	12	0	0	0
	fad	0	0	0	9	4	7
	gla	0	0	0	3	7	4
	mas	0	3	1	2	5	5

Figure 7: Cross validation result as confusion matrix.

3. FUTURE WORK:

It would be interesting the implement a robust LDA method similar to robust PCA which uses minimum covariance determinant (MCD) as the scatter functional for the covariance matrix. Following the same lines it would be interesting to find better LDA rule for classification task using a kind of cross-validation technique where for each subset of data is used for finding the LDA rule and it is then tested on the validation set. Then the rule which best performs on the validation set is selected and then the accuracy on the test dataset is computed.

4. DISCUSSION and CONCLUSION:

It can be concluded that the classification task on a real world dataset is daunting task with the naïve application of algorithms would hardly work but the linear discriminant analysis works pretty well on the data given here with a accuracy of 75% on the test data. As seen from the scatter plot that most of the groups could not be separated, that could be one of the reason for the reduced accuracy.

APPENDIX

A.1 Correlation Coefficients for the data

Below figure shows the correlation values for the data. The attributes with highest correlation have been highlighted.

	I0	PA500	HFS	DA	Area	A.DA	Max.IP	DR	P
I0	1	-0.39364731	0.02845465	0.81960638	0.56009778	0.6120695	0.82366786	0.73325217	0.9886973
PA500	-0.39364731	1	0.50901926	-0.08981716	0.08354689	0.2298367	-0.05040076	-0.07705354	-0.3457152
HFS	0.02845465	0.50901926	1	0.1069767	0.20605871	0.3560277	0.37082738	0.0115922	0.1023619
DA	0.81960638	-0.08981716	0.1069767	1	0.73113235	0.6483337	0.75322651	0.97420242	0.7740281
Area	0.56009778	0.08354689	0.20605871	0.73113235	1	0.8301723	0.73525781	0.67581022	0.5740731
A.DA	0.61206951	0.22983666	0.35602765	0.64833369	0.8301723	1	0.81281518	0.54069504	0.679363
Max.IP	0.82366786	-0.05040076	0.37082738	0.75322651	0.73525781	0.8128152	1	0.60028988	0.8618369
DR	0.73325217	-0.07705354	0.0115922	0.97420242	0.67581022	0.540695	0.60028988	1	0.6659872
P	0.98869732	-0.34571522	0.10236192	0.77402811	0.57407314	0.679363	0.86183691	0.66598724	1

REFERENCE

[1] J Jossinet (2010). UCI Machine Learning Repository [http://http://archive.ics.uci.edu/ml/datasets/Breast+Tissue]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Jossinet J (1996) Variability of impedivity in normal and pathological breast tissue. Med. & Biol. Eng. & Comput, 34: 346-350.

[3] Silva JE, Marques de Sá JP, Jossinet J (2000) Classification of Breast Tissue by Electrical Impedance Spectroscopy. Med & Bio Eng & Computing, 38:26-30.