Midterm Report

CS-489 | Dr. Wang

March 21, 2024

Casey Lang

Table of Contents

1. Project Outline

This project is intended to work as a learning tool for data extraction from images. The project uses a data set of 160 prostate histopathology images, some of which are of cancerous areas and some of which are not, which was collected by Wouter Bulten. As a case study, the data set is configured in R for variable extraction and image classification. Then, variable extraction and image classification is performed, followed by sensitivity and specificity analysis. This process will then be displayed and explained on a website. This website will explain to the user how to undergo the process of extracting data from images, including informative images and links to helpful sources to allow the user to expand upon the information provided by the website. Additionally, an interactive tool will be included on the website that allows the user to upload an image and extract variables from it, providing them with a hands on introduction to variable extraction.

2. Case Study

a. Data Selection

Wouter Bulten's prostate histopathology image data set was chosen as an adequate data set for variable extraction and image classification due to its inclusion of whether or not an image corresponds to a cancerous region of the patient's prostate. This allows for classification methods to be tested easily on the data set. Since the cancerous images are known, the accuracy of predictions of cancer in the image based on extracted variables can easily be tested.

b. Data Configuration

The data set is provided as a CSV file. When uploaded to R, the data set is converted to a dataframe with 4 variables, the id of the image, the slide that the image is found on, whether or not the image contains cancer, and the id of an image in its respective slide.

This dataframe is shown in Figure 1.1.

```r
df <- read.csv("/Users/Owner/Downloads/CS489/peso_testset_mapping.csv")
df
```

Description: df [160 × 4]

id <int>	slide <int>	type <chr>	annotation_id <int>
1	1	non-cancer	0
2	1	non-cancer	1
3	1	cancer	2
4	1	non-cancer	3
5	5	non-cancer	0
6	5	non-cancer	1
7	5	cancer	2
8	5	cancer	3
9	7	non-cancer	0
10	7	non-cancer	1

1-10 of 160 rows

Previous 1 2 3 4 5 6 … 16 Next

Figure 1.1

The cancer classification variable is stored as a character variable in the dataframe. For classification, this variable needs to be stored as a numerical variable. This is done in Figure 1.2, where the mutate() command is used to create a new column, “Cancer,” that stores a 0 for cancerous images and a 1 for non-cancerous images.

```r
dfBinary <- df %>% mutate(Cancer=ifelse(type=="cancer",1,0))
dfBinary
```

Description: df [160 × 5]

id <int>	slide <int>	type <chr>	annotation_id <int>	Cancer <dbl>
1	1	non-cancer	0	0
2	1	non-cancer	1	0
3	1	cancer	2	1
4	1	non-cancer	3	0
5	5	non-cancer	0	0
6	5	non-cancer	1	0
7	5	cancer	2	1
8	5	cancer	3	1
9	7	non-cancer	0	0
10	7	non-cancer	1	0

1-10 of 160 rows

Previous 1 2 3 4 5 6 … 16 Next

Figure 1.2

c. Variable Extraction

The code for variable extraction is shown below in Figure 1.3.

```
21 ```{r pressure, echo=FALSE}
22 darkList <- c()
23 lightList <- c()
24 redMeanList <- c()
25 greenMeanList <- c()
26 blueMeanList <- c()
27 redSDList <- c()
28 greenSDList <- c()
29 blueSDList <- c()
30 for (k in 1:160){
31 newImage <- load.image(paste("/Users/Owner/Downloads/CS489/peso_testset_png/",k,".jpg", sep=""))
32
33 redMeanList[k]<-mean(as.numeric(R(newImage)))
34 greenMeanList[k]<-mean(as.numeric(G(newImage)))
35 blueMeanList[k]<-mean(as.numeric(B(newImage)))
36
37 redSDList[k]<-sd(as.numeric(R(newImage)))
38 greenSDList[k]<-sd(as.numeric(G(newImage)))
39 blueSDList[k]<-sd(as.numeric(B(newImage)))
40
41 grayImage <- grayscale(newImage)
42 darkPixels<-grayImage<0.5
43
44
45 lightPixels<-grayImage>0.9
46
47
48 darkList[k] <- sum(darkPixels) / (2500*2500)
49 lightList[k] <- sum(lightPixels) / (2500*2500)
50 }
51 ```
```

Figure 1.3

First, empty vectors are created that will eventually contain the variables. Then, a for loop is used to extract variables from each of the 160 images. On line 31, a new image is uploaded as a variable in R. Next, the as.numeric(R()), as.numeric(G()), and as.numeric(B()) commands are used to extract the red, green, and blue channels from each pixel in the images as numerical values. This allows the means and standard deviations for the color channel values of the pixels in each image to be calculated and used as representations of the red, green, and blue intensities in the images. On line 41, a grayscale copy of the image is created. This allows for a simple threshold to be used to obtain counts of

dark and light pixels in the grayscale image. The proportions of light and dark pixels in an image are then calculated by dividing the counts by the number of pixels in an image. All of these extracted variables are stored in their respective vectors that were created at the beginning of Figure 1.3.

These variables are then added into the dataframe using the mutate() function. This is shown in Figure 1.4, along with the resulting dataframe.

```r
```{r}
dfFinal =dfBinary %>% mutate(dark_proportion=darkList,light_proportion=lightList,dark_light_ratio=darkList/lightList,red_mean=redMeanList,green_mean=greenMeanList,blue_mean=blueMeanList,red_sd=redSDList,green_sd=greenSDList,blue_sd=blueSDList)
dfFinal
```
```

Description: df [160 × 14]

id <int>	slide <int>	type <chr>	annotation_id <int>	Cancer <dbl>	dark_proportion <dbl>	light_proportion <dbl>	dark_light_ratio <dbl>	red_mean <dbl>
1	1	non-cancer	0	0	0.09275136	0.24594832	0.37711727	0.8390670
2	1	non-cancer	1	0	0.13301120	0.38376704	0.34659360	0.8294967
3	1	cancer	2	1	0.11358992	0.19444960	0.58416124	0.8206209
4	1	non-cancer	3	0	0.12447504	0.18645680	0.66758112	0.8028730
5	5	non-cancer	0	0	0.10015632	0.16084416	0.62269168	0.8330582
6	5	non-cancer	1	0	0.15073184	0.29130976	0.51742805	0.8128536
7	5	cancer	2	1	0.16189360	0.12642176	1.28058334	0.7697798
8	5	cancer	3	1	0.20591024	0.07046448	2.92218491	0.7492862
9	7	non-cancer	0	0	0.14777104	0.15195056	0.97249421	0.8048996
10	7	non-cancer	1	0	0.11614064	0.20761168	0.55941284	0.8187197

1-10 of 160 rows | 1-9 of 14 columns

Previous 1 2 3 4 5 6 … 16 Next

Figure 1.4

d. Classification Technique

For a classification technique, logistic regression is used as an example. This is done using the glm() function, as shown in Figure 1.5.

```r
fit_logistic<-glm(Cancer~red_mean+green_mean+blue_mean+red_sd+green_sd+blue_sd+dark_proportion+light_proportion+dark_light_ratio,data=dfFinal,family = binomial(link="logit"))
summary(fit_logistic)
```

```
Call:
glm(formula = Cancer ~ red_mean + green_mean + blue_mean + red_sd +
    green_sd + blue_sd + dark_proportion + light_proportion +
    dark_light_ratio, family = binomial(link = "logit"), data = dfFinal)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -46.876     28.001  -1.674 0.094112 .
red_mean         -114.239     28.851  -3.960 7.51e-05 ***
green_mean       -288.173     79.093  -3.643 0.000269 ***
blue_mean         444.447     97.237   4.571 4.86e-06 ***
red_sd           -105.516     46.216  -2.283 0.022424 *
green_sd         -131.984    104.383  -1.264 0.206081
blue_sd           262.383    114.936   2.283 0.022438 *
dark_proportion   -10.061     17.175  -0.586 0.557996
light_proportion    1.148     14.114   0.081 0.935150
dark_light_ratio    1.423      1.626   0.876 0.381287
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 219.78  on 159  degrees of freedom
Residual deviance: 116.57  on 150  degrees of freedom
AIC: 136.57

Number of Fisher Scoring iterations: 7
```

Figure 1.5

Then, a confusion matrix, shown in Figure 1.6, is created to show correct and incorrect classifications based on the logistic regression model.

```r
library(InformationValue)
InformationValue::confusionMatrix(actuals = dfFinal$Cancer,predictedScores = predict(fit_logistic))
```

Description: df [2 × 2]

	0 <int>	1 <int>
0	82	20
1	7	51

2 rows

Figure 1.6

The model correctly predicts 82 non-cancerous images out of 89 total non-cancerous images and 51 cancerous images out of 71 total

cancerous images. Finally, sensitivity (the true positive rate) and specificity (the true negative rate) are calculated in Figure 1.7.

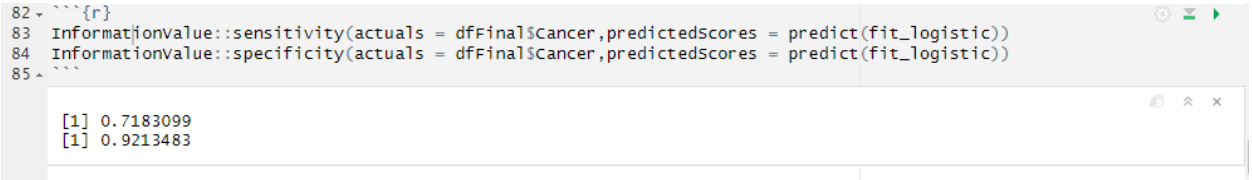

Figure 1.7

A sensitivity of 0.718 and a specificity of 0.921 are obtained. This shows that the logistic regression model is extremely specific, but only moderately sensitive. It can be concluded that further variable extraction and exploration of classification techniques is needed to obtain adequate sensitivity. Links to resources explaining other potential techniques will be included on the website.

3. Next Steps

Moving forward, the focus of this project will shift towards the creation of a website that will act as a learning tool for image variable extraction. The case study will be used as an example, showing the steps to properly perform variable extraction and image classification. Sources will be found and included in the website that will provide the user with resources that will allow them to go beyond the more introductory approaches outlined on the website. Additionally, a page on the website will be created that takes an image file as an input and outputs the 8 variables that were extracted in the case study.

4. Conclusion

The majority of the project so far has been completion of the case study. This provides an adequate example of variable extraction in practice that can be explained effectively on the learning tool website. The remainder of the project will consist mostly of web design, using HTML, PhP, and CSS to create an interesting and engaging learning tool.

5. References and other Resources

References

Bulten, Wouter, Péter Bándi, Jeffrey Hoven, Rob van de Loo, Johannes Lotz, Nick Weiss, Jeroen van der Laak, Bram van Ginneken, Christina Hulsbergen-van de Kaa, and Geert Litjens. “Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard.” Scientific reports 9, no. 1 (2019): 1-10.

Bulten, W., Bándi, P., Hoven, J., Loo, R. van . de ., Lotz, J., Weiss, N., Laak, J. van . der ., Ginneken, B. van ., Hulsbergen-van de Kaa, C., & Litjens, G. (2021). PESO: Prostate Epithelium Segmentation on H&E-stained prostatectomy whole slide images (1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5137717

Other Resources

- https://cran.r-project.org/package=dplyr
- https://CRAN.R-project.org/package=imager
- https://rdocumentation.org/packages/InformationValue/versions/1.2.3
- https://stackoverflow.com/