CS4780/CS6780 Final Project

Funshang Wu, Mokter Hossain, and Jonathan Pinder
CS 4780/6780 FINAL PROJECT

***Abstract*—A serious issue in healthcare today is the increasing resistance of microbes to antibiotics. We are seeking to investigate whether the physical appearance of microbes correlates in anyway with their tolerance to antibiotics. Our findings hope to further research into evolving medical practices to fight rapidly evolving microbes. This paper studies microbial resistance to carbenicillin and tobramycin.**

I. INTRODUCTION

Bacteria can be classified as Prokaryotic and Eukaryotik. Although these organisms are relatively small they have big impacts in our lives. The phenotype of a bacteria tells us how the organism acts. In this paper phenotype will refer to the resistance to antibiotics **carbenicillin** and **tobramycin**. The morphotype we are referring to is the bacteria *Pseudomonas aeruginosa*. Unfortunately, this bacteria is opportunistic: disrupts mucous membrance/skin, affects cancer therapy and burn wound patients. *Pseudomonas aeruginosa* is very diverse and is resistant to many of today's antibiotics.

We were determined to do this project mostly to do the code for this project as it will be an interesting topic to code up. We decided this topic as it involved extensive research on the topic as well as extensive coding were needed to solve the task. Neural network is that type of topic for each of us and gave us a chance to code something that is not only difficult, but also took some long time to complete its execution.

II. EVALUATION

Pseudomonas aeruginosa could be found in various environmental habitats like animal, human that is one of the major opportunistic pathogens that is able to cause soft tissue infection on the weak hosts. Bacteria can inherent and evolve resistance to antibiotics. So, our goal is to analyze the morphotypes and its corresponding predict label to conclude a model that can take any morphotype, then to predict its label, which is numerical way is the degree of resistance to our antibiotics. Approximately, 90+ morphotype images, whereas that Carb.lag.delta define as the degree of antibiotic resistance carbenicillin antibiotic, Toby.lag.delta define as the degree of antibiotic resistance tobramycin antibiotic. Numbers are meaningless without any further interpretation; however, what we can do is to cluster them and build model around each cluster. We applied a number of approaches: such as linear regression, logistic regression, and neuron network to justify which performs higher accuracy.

III. METHODOLOGY

Our core methodlogy for this experiment was using a linear regression model. The following methodologies were used:

- Phenotypic Analysis
- Phenotype Classification
- Morphotype Recognition
- Model Training
- Testing

Once completed, we used the K-Means++ algorithm to ascertain our findings. K-Means++ allowed us to find patters in our analysis of the phenotype and test the validity of our models.

IV. DATA

The data used in the research came from a research lab at Georgia Tech, United States. Our data set consists of a variety of images, similar to Figure 1. Along with a spreadsheet with vectorized numerical values of the images. The spreadsheet's information helps to discern the important aspects and features and dismiss any "noisy" data we were given. The dataset comprised of pictures, and a spreadsheet containing numerical values of each pictures. The spreadsheet helps us gain more understanding on the more significant data objects and allows to dismiss the noisy data. See Figure:

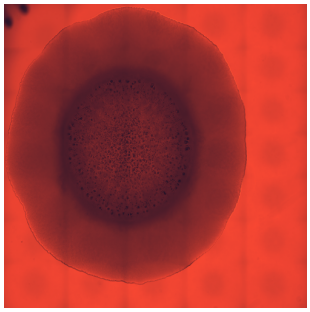

Fig. 1: A sample of Pseudomonas aeruginosa

V. METHODS

The methods used came from taking 10 random samples at a time of Pseudomonas aeruginosa. Our samples had different tolerances to offset any biases in our results and model creation.

For the Linear Regression Model, we solved for beta with the following equation:

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}. \tag{1}$$

Using the Linear Regression Equation:

Linear Regression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{2}$$

We got the following test results.

```
>>>
==== RESTART: C:\Users\sheng\Desktop\DataScience\finalProj\DS_FP_LinReg.py ====
Column headings: Index(['carben', 'tobra', 'N1', 'N2', 'P1', '1-P1', 'LogY1', 'LogY2'], dtype='object')
Beta1: [[-0.02991294]
 [-0.02127896]
 [-0.010084  ]
 ...
 [ 0.00708706]
 [ 0.0366846 ]
 [ 0.01044843]] (12544, 1)
Beta2: [[ 1.08887546e-02]
 [ 2.92319086e-03]
 [-3.32020889e-05]
 ...
 [ 1.22466013e-03]
 [-1.36064299e-02]
 [-9.26335136e-04]] (12544, 1)

test_image: [[250. 143.  61. ... 119.  59. 255.]] (1, 12544)

predict_value_Y1 carbenicillin resistance: [[25.00060737]] ,    predict_value_Y2 tobramycin resistance: [[2.99981173]]

actual_value_Y1 carbenicillin resistance: 21.0 ,        actual_value_Y2 tobramycin resistance: 1.0

accuracy : [[-19.05051127]] %, [[-199.9811728]] %
>>>
```

Fig. 2: Linear Regression Test Results

The second method we used was logistic regression: We wanted to try another algorithm to classify our results and see how much more accurate of a prediction we could receive. Logistic Regression:

$$p = \frac{1}{1 + e^{-\beta^T x}} \tag{3}$$

After the Logistic Regression model was applied. The results indicates that ..see Figure 3

```
>>>
==== RESTART: C:/Users/sheng/Desktop/DataScience/finalProj/DS_FP_LogReg.py ====
Column: Index(['carben', 'tobra', 'N1', 'N2', 'P1', '1-P1', 'LogY1', 'LogY2'], dtype='object')
Iteration 2000 ,
X value is [[1.28627451 1.0        1.83137255 ... 1.89803922 1.84705882 1.        ]
 [1.01960784 1.43921569 1.76078431 ... 1.53333333 1.76862745 1.        ]
 [1.01176471 1.19607843 1.71372549 ... 1.2627451  1.72156863 1.        ]
 ...
 [1.         1.59607843 1.79215686 ... 1.60784314 1.79607843 1.        ]
 [1.01960784 1.11764706 1.72156863 ... 1.25490196 1.7254902  1.        ]
 [1.00392157 1.         1.25882353 ... 1.         1.28235294 1.        ]]

predict image:  0       value:  0.654600799495498
predict image:  1       value:  0.508576693061237
predict image:  2       value:  0.5387045013530832
predict image:  3       value:  0.9509321667404974
predict image:  4       value:  0.8324564559581593
predict image:  5       value:  0.7904852612923182
predict image:  6       value:  0.7930297594044247
predict image:  7       value:  0.95028648139183l
predict image:  8       value:  0.7797722900991857
predict image:  9       value:  0.5823636494791448
MAX predict carbenicillin resistance: 0.9509321667404974 ,      MAX predict tobramycin resistance: 0.04906783325950259

actual_value_Y1 carbenicillin resistance: 1.0 , actual_value_Y2 tobramycin resistance: 0.0

accuracy : 95.09321667404974 %,  -4.90678332595026 %
>>>
```

Fig. 3: Logistic Regression Test Results

Then we used the Elbow Method mode, 4l. The results indicates that six clusters should be appropriate in order to represent the data..

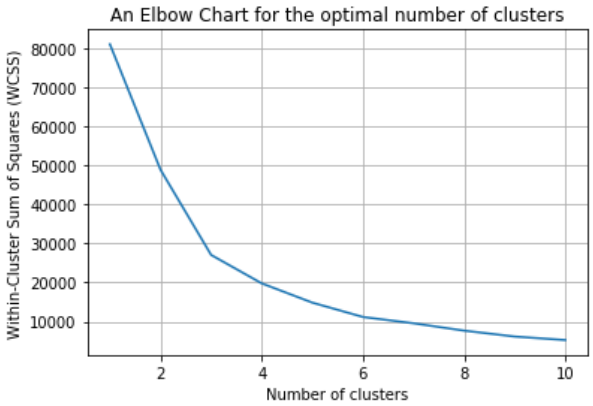

Fig. 4: Elbo Method Result

Elbo Method model was applied. The results indecates that six clusters should be appropriate in order to represent the data. See figure 4.

We also used K-Means++ clustering algorithm to ascertain our findings. See figure 5

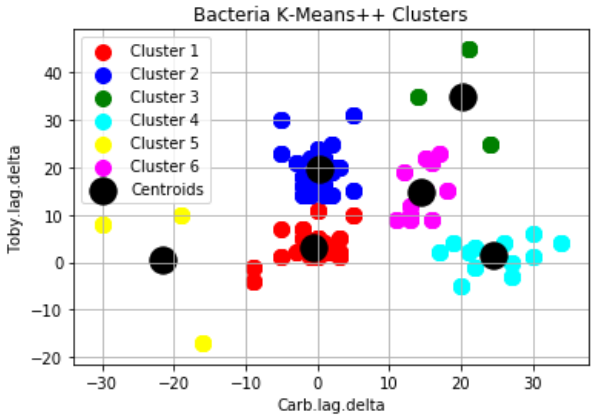

Fig. 5: KMeans++ Method Test Result

VI. Experiments

There was a test image that we used to predict the values. Using our Linear Regression model, we get the values shown in I.

Image	Beta1(carben)	Beta2(tobra)
1	-0.02991294	1.08887546e-02
2	-0.02127896	2.92319086e-03
⋮	⋮	⋮
10	0.01044843	-0.26335136e-04

TABLE I: Linear Reg Results

Using our model we tested our models efficiency on our test image and got II

TestImage	Carbenicillin Resistance	Tobramycin
Predicted	25.00060737	2.99981173
Actual	21.0	1.0
Accuracy	-19.05051127%	-199.9811725%

TABLE II: Linear Reg Results

Method	Antibiotic 1	Antibiotic 2
Linear Regression	**0.12**	0.18
Logarithmic Regression	0.16	**0.13**

TABLE III: Predictions Mean Squared Error.

Using the Logistic Regression Model we received the above Summarized Results III.

VII. Discussion

Even though, we only fed ten pictures at a time to find its Beta value, our prediction on new image was quite accurate. Based on the experience we had from this project, if one needs

to have a perfect model, it must meet several requirements. First, the picture needs quality with high resolution that allows every pixel to be distinguishable. This would help the accuracy improve. Second, computing power, when we computed, we resized the picture to 64x64 and it is float64, because some of the Python numpy libraries could not solve float32 with some matrix operations, that means one picture is 32768 bytes. So, 10x(64x64) was our training data size. Especially, when we performed dot product operation of Nx10 * 10xN, it takes at least 5 minutes just to finish one step. The best PC used was on 4.3Ghz processor and 16G RAM. Third, the logistic regression algorithm takes lots more time to find its pattern. During the maximum likelihood, we must run at 10 x (64x64) + 10 x (64x64) * learning steps which is unknown basing on the value of pixels. The Gradient Ascent is similar to the Restricted Boltzmann Machine we tried has very long-time training. One small error would have to start all over very time consuming.We tried to use a neural network to train our data to no avail, the code is provided.

VIII. Conclusion

The linear regression algorithm was very good at determining the phenotype of the carbenicillin antibiotic resistance within 20% accuracy. However, the accuracy of determining the tobramycin antibiotic resistance was much less accurate, missing the actual value by about 200%.

The logistic regression algorithm overall was probably the better algorithm, the carbenicillin's resistance accuracy was about 95% and the tobramycin's resistance was around 4%.

References

[1] Sutskever, Ilya , Hinton, Geoffrey; and Taylor, Graham. *The Recurrent Temporal Restricted Boltzmann Machine*, Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008.