Bacteria Classification

CSC 6780 FINAL PROJECT

Kevin Horning
Brent Mckinney
Leonard Sanders

***Abstract*—The goal of our project is to obtain insight from a data set consisting of images of bacteria. We used several algorithms and techniques in attempts classify images as antibiotic resistant or not.**

I. PREPROCESSING IMAGES

Our first goal was to determine how to process the images. In order for us to perform any sort of calculations, we needed to get the images in a mathematical format so that we could feed our algorithms.

1) Downsize each image
2) Convert each image to grayscale
3) Convert each image to a vector.

Each image can be understood as a matrix of RBG values. Knowing this, we decided to vectorize each of these matrices. This way we could use the vectors in our algorithms. Before turning each one into a vector, we resized the image to smaller dimensions, and converted them to grayscale. The reason behind this was due to the size of the images. In order to spend more time working on solutions, we needed to scale down the size of our data.

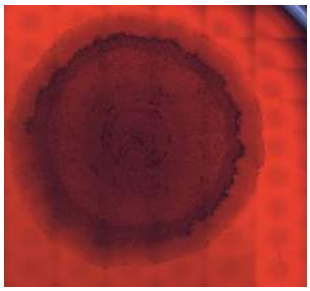

Fig. 1: Original Bacteria Image

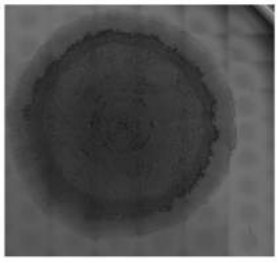

Fig. 2: Grayscale Bacteria Image

At this point we have a csv file containing a row for each image (Grayscaled and Resized).

Image	rgb value	rgb value	rgb value	...	rgb value
PIL-1_3dayLBCR-1	101	104	105	...	92
PIL-2_3dayLBCR-1	103	105	106	...	100
⋮	⋮	⋮	⋮	⋮	⋮
PIL-344_3dayLBCR-4	178	181	182	...	166

TABLE I: Here's a look into what our image data looks like.

II. NORMALIZING OUR DATA

The next thing we needed to do was normalize our training data. The purpose behind this being to reduce redundancy in our data, and have it all on a common scale. We were able to create a program that outputs a csv file containing new columns of our training data, all between the values 0 and 1. Along with these values, we included columns stating whether that image was resistant to antibiotic 1 and 2, based on a threshold of 30%.

Image	Res 1	Res 2	Res to 1?	Res to 2?
PIL-1_3dayLBCR-1	0.776	0.225	Yes	No
PIL-2_3dayLBCR-1	0.347	0.475	No	No
⋮	⋮	⋮		
PIL-344_3dayLBCR-4	0.408	0.675	No	Yes

TABLE II: Here's a look into what our normalized data looks like.

III. EXPERIMENTS

The first approach we used to determine antibiotic resistance was K-Nearest Neighbors (KNN) implementation. The feature vectors of KNN algorithm were the vectorized bacteria images. We calculated the distance of the unlabeled bacteria sample to all other bacteria samples with the L2 norm. The training data vector that had the smallest distance was called the 'nearest neighbor' of the unlabeled bacteria vector because it has the features that are most similar to the unlabeled vector features.

The resistance to both antibiotics of this nearest neighbor was considered the best estimation of the resistivity of the unlabeled bacteria.A test of this algorithm's effectiveness was run by removing 10 data vectors from the training data set and using them as input (test data).

IV. Conclusions

It was found that our KNN algorithm identified nearest neighbors for each of the vectors that had the same resistivity labels 70% of the time. In other words, based on our sample with 10 bacteria removed from the training data and used as input, the accuracy of this algorithm is 70%. We can see below that it correctly predicted the resistance 70% of the time.

Bacteria sample name	Antibiotic Resistance Actual 1	Antibiotic Resistance Actual 2	Antibiotic Resistance Predicted 1	Antibiotic Resistance Predicted 2
PIL-313 3dayLBCR-4	No	No	Yes	No
PIL-317_3dayLBCR-4	No	No	Yes	No
PIL-329 3dayLBCR-2	No	No	No	No
PIL-331 3dayLBCR-4	No	Yes	No	No
PIL-333_3dayLBCR-4	No	Yes	No	Yes
PIL-334 3dayLBCR-4	No	No	No	Yes
PIL-337 3dayLBCR-4	No	Yes	No	Yes
PIL-338_3dayLBCR-4	No	No	No	Yes
PIL-340 3dayLBCR-4	No	Yes	No	No
PIL-344 3dayLBCR-4	No	No	No	No

Fig. 3: KNN results

We also tried logistic regression, as our hope would be to use each image vector $\mathbf{x}$ as input, but were unsuccessful in obtaining a $\boldsymbol{\beta}$ that would give us any degree of accuracy. Using this method, we would be able to determine if each image was resistant or not based on the formula alone, as the result of logistic regression is binary.