Riley's CS251 Final Project

Abstract

For this project, I used analysis techniques such as KNN classification and K-means clustering, and a GUI as a visualization aid, to plot and analyze a systems bio data set. The programs work with data from CSV (comma separated values) files whose first row contains the header names for the columns, the second row specifies the type of data in each column, and the following rows contain the data values. Previously created python files including data.py, PCAData.py, view.py, display.py and analysis.py were used to plot and analyze the data points. I used my classify.py file to classify the data, and I used the k-means clustering feature of my GUI to cluster and plot the bioluminescence data in the data set. I also added a feature to my GUI that allows the user to view the changes in bioluminescence of the cells over time. The bioluminescence data was collected from 123 cells from a slice of a mouse brain, in a 2009 experiment.

Problem

The bioluminescence data was originally collected in 2009 during an experiment by Webb, Angelo, Huettner, Herzog, and PNAS. the bioluminescence data was recorded in three conditions: the base condition (sampled one hour apart for 144 hours), after treating the cells with a toxin (sampled one hour apart for 144 hours), and after washing the toxin out (sampled one hour apart for 142 hours). The data was discussed in more detail in 2012 by Webb, Taylor, Thoroughman, Doyle III, Herzog, and PLoS Computational Biology. The systems bio data file that I used contains position data, and the bioluminescence data at each hour in each condition. The main questions that I set out to answer by analyzing this data set are the following:

1.) Are tightly-synchronized cells also close together spatially?

2.) Does the environment of the cells affect their relationship with each other?

3.) Does the bioluminescence of each cell change at the same rate?

Methods

To solve the first problem, I used my classify.py file to classify the data using a KNN classifier. I found the categories for the training and testing sets by conducting a k-means cluster analysis on the entire data set, and saved the category labels as a column at the end of the original data set. Then I divided the data randomly to crate a training set with 61 cells (data points), and a testing set with 62 cells. I then classified the data with the KNN classifier (using three nearest neighbors in the distance calculation), and saved the resulting categories of the testing set to knnclassifythree.csv. Then I plotted the data points, with the x-position on the x-axis and the y-position on the y-axis, and with a different color representing each category.

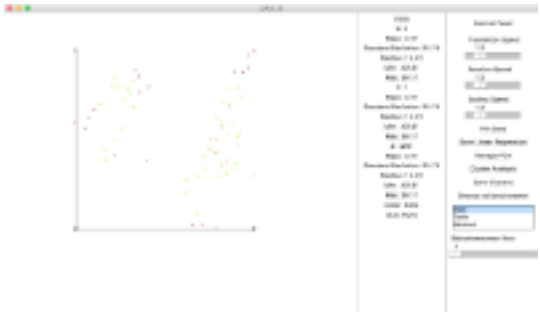
To solve the second problem, I performed three k-means cluster analyses, and plotted the points with different colors for each cluster. In all three analyses, I used k = 3, and plotted the x-position on the x-axis and the y-position on the y-axis. The first analysis clustered the points based on the bioluminescence data taken when the cells were in the base condition. The second analysis clustered the points based on the data taken after a toxin had been added to the cells. The final analysis clustered the points based on the data taken after the cells had been washed. I then compared the three plots and noted any differences or similarities. To make it easier to select a large block of column headers to use for clustering, I added a select() method to my SelectClusterHeaders class in display.py, so that if you right click, then all of the points between the last two that were selected, will then be activated, so you don't have to click on each header individually.

To answer my final question, I added a listbox and a slider to my GUI, which allow the user to view plot the bioluminescence data, and view how the relative differences between the bioluminescence of each cell changes over time, in each experimental environment. The ListBox is where the user can select the environment, and the slider is where the user can choose which hour to look at (since data was taken each hour for 144 hours or 142 hours, depending on the environment). I added the ListBox and Slider to the control panel when building the controls. I added a boolean field to the DisplayApp class, called self.circadian, which was set to true if the systems bio data set was opened, and was false otherwise, since these features only work on that data set. I also added an updateBio() method to the DisplayApp class, which gets the selections from the ListBox and slider, and changes the colors of all of the data points, based on the column of bioluminescence data in the specified environment at the given hour. The colors of the points are based on a blue-yellow gradient, where blue represents the point with the lowest bioluminescence, and yellow represents the point with the highest bioluminescence.

Results

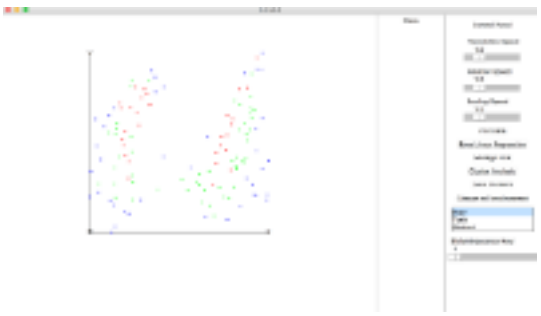
The confusion matrix resulting from my KNN classification of the data in my first experiment, as well as the plot of the data, with the colors given by the calculated categories, can be seen below. We can see that the points of the same category are mostly located neat each other, however not all of the points of each category are in the same area. There were small clusters of each category located throughout the slice of cells.

(Figure 1)

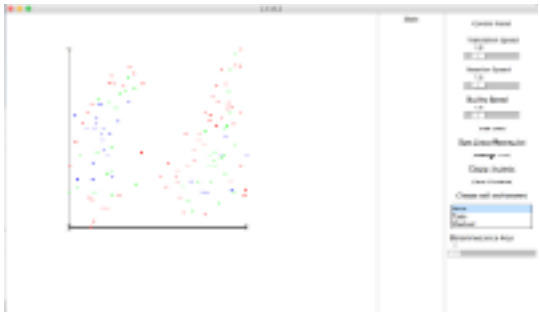


(Figure2)

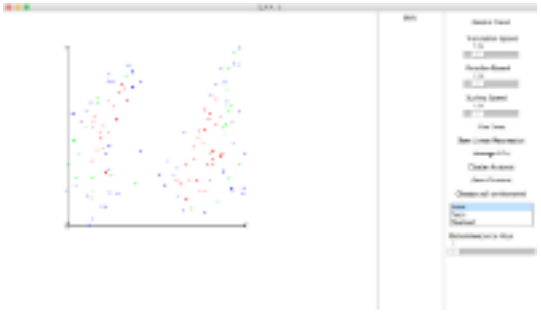
The plots of the three k-means cluster analyses that I conducted in my second experiment can be seen in the three images below. The first image is of the analysis based on the data from cells in the base condition, the middle image is of the analysis based on data from cells treated with a toxin, and the final image is of the analysis based on data from cells after being washed. Looking at the plots, we see that there are barely any differences between the first and second plots, and there are only slight differences in the third plot compared to the first two plots. These differences could have been caused by the randomness of the initial step of choosing random means during the k-means analyses, r maybe by some other kind of data collection error or inconsistency. The lack of major differences in the plots indicates that the cells mostly remain in the same clusters, no matter which environment they are put in.



(Figure 3)

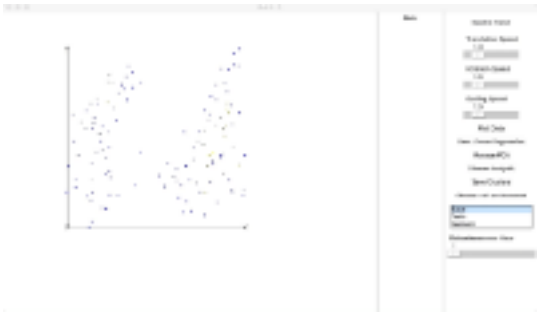


(Figure 4)



(Figure 5)

A GIF showing the color changes as the slider moves, for each experimental condition, can be seen below. The colors of the points are based on a blue-yellow gradient, where blue represents the point with the lowest bioluminescence, and yellow represents the point with the highest bioluminescence. Moving the slider from 1 to 144, for all three environments, we see that the cells on the outer edges of the two separate sections of cells don't change color very much, indicating that their bioluminescence changes minimally compared to the amount that the bioluminescence changes in the cells closer to the center of the two sections of cells.



(Figure 6)

Conclusion

During this project, I learned how to design and conduct an experiment from start to finish, including picking a data set and questions to answer, designing experiments to find answers to the questions, conducting the experiments, and analyzing the results. I learned how to do all of this using computational tools and algorithms. The main results from my experiments were that 1.) closely-synchronized cells tend to be relatively near each other spatially, 2.) the environment of the cells doesn't effect the relationships between the cells, and 3.) the bioluminescence of the cells on the edges of the two sections change less over time than the bioluminescence of the cells towards the center of the sections.