Daniel Lee
18379784
Assignment 2

Connectomics

Summary

Connectomics is an attempt to map out all the connections in the brain. However, current technology presents limitations to do this accurately. In an attempt to map out all the neuronal connections, kaggle hosted a competition for data scientists to attempt to predict the direct connection between neurons based on fluorescent calcium indicators and location of the neurons. My attempt in predicting connections between two neurons revealed that it is important to process the data effectively. It is important to remove noise in the fluorescence levels. It is also important to remove the global activity of neurons, which is when all the neurons experience a spike in fluorescence levels. Removing global activity of neurons is important because global activity can interfere with determining presence of a direction connection between two neurons. Because of global activity, it may seem like two neurons are connected even when they are actually not connected. My attempt does not remove the global activity because I was not successfully able to implement a method to eliminate a lot of the global activity. That is perhaps the single biggest reason why my method seems to be ineffective. For the training data, I used a dataset containing 100 neurons, which contains 179,497 fluorescence fluorescent calcium indicators measured at 20 millisecond time intervals for each neuron. I processed the data by calculating the differences between any two consecutive time intervals for each neuron. Then, I zeroed all the times with small differences, calculated the correlation matrix among all the neurons, and then determined presence of a direct connection based on the correlation coefficient between two neurons. For my testing data, I used a different dataset containing 100 neurons with 179,497 fluorescence levels as well. I analyzed my data by calculating the true negatives, true positives, false positives, false negatives, sensitivity, precision, and false positive rate.

Competition Question

Given a network of neurons, predict whether two neurons are directly connected.

Data Processing

The raw data is a time series of the activity of 100 neurons. For each neuron, the data has 179,497 fluorescence fluorescent calcium indicators measured at 20 millisecond time intervals. Fluorescent calcium indicators indicate presence of neural activity. Since fluorescent calcium indicators generate a lot of noise, I decided to reduce the noise using the difference method. That is, I calculated the difference between any two consecutive time steps. Then, I zeroed the differences that were small to eliminate a lot of the noise using a specific threshold level. This was done for all the neurons. Figure 1 shows plots of fluorescence levels for each step of the data processing for one neuron for 50 time intervals.

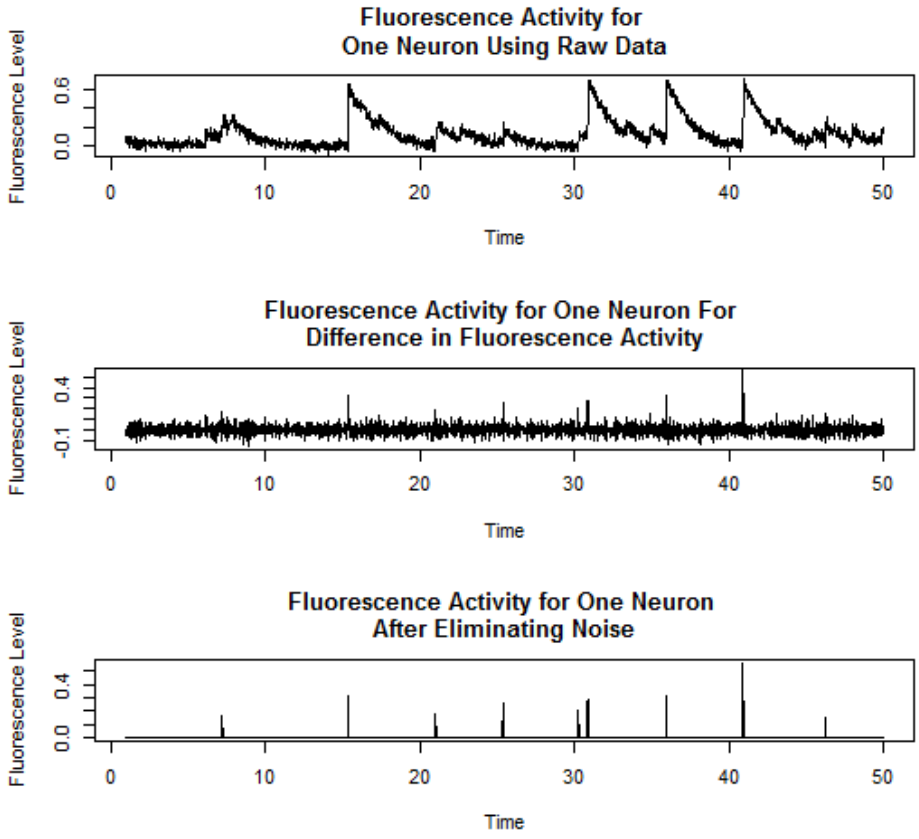
Data Analysis

Summary of Winning Method

The team who performed the best in their prediction used partial correlation to determine whether there's a connection between two neurons. This team first processed the data based on the signals given by the fluorescence signals. There are two problems with the raw fluorescence signal data. First is that the raw fluorescence signal is noisy. Second is that there is global neuronal activity in which all the neurons will demonstrate high fluorescence signal. This winning team filters out the fluorescence signals so that the noise and the global neuronal activity are filtered out. After that, the team calculates the partial correlation among all the neurons based on the filtered data and determine which neurons are connected. The team then evaluated their performance using AUROC and AUPRC. They compared this partial correlation method to other methods including the following: standard

(Pearson) correlation, generalized transfer entropy, and GENIE3. Among all the different methods, the partial correlation method performed the best.

Figure 1: The three plots reveal data processed by me. The very top plot represents raw fluorescent levels for one neuron. The middle plots the difference between any two consecutive time steps for the same neuron. The last plot represents fluorescence activity for one neuron after eliminating noise.



Another winning team used two methods to predict the presence of connections between neurons. First was to look at the difference between any two consecutive time steps. The spike differences smaller than a certain threshold were set to zero to eliminate background noise. They tried different threshold levels and evaluated the performance of their predictions. The second method was using the sequential Monte Carlo framework that finds the probability of the neuron spiking in each time step. After that, this team removed all the spikes that were part of the synchronization rate. Then, they calculated the correlation of the fluorescence levels among all the neurons. They determined the presence of a connection between neurons based on the correlation coefficients between two neurons. Then they used a network deconvolution algorithm to determine more accurately the presence of connections between two neurons. The performance for both methods were evaluated using the AUC. This team performed fourth overall in the public leaderboard for kaggle.

Another team, who ranked 5th in the competition, used the Random Forest classifier to predict connections between two neurons. The first step in their method was to preprocess the raw data in order to retrieve spike times of neurons. Then, they used several different base predictors, which were cross-correlation, cross-correlation with a lag of one frame, Generalized Transfer Entropy, and Information Gain. Then they used Random Forest to make classifications. One thing they did for the Random Forest method was to include topological information to determine the direction of the neuron pathway. The performance of this method was evaluated using ROC curve.

My Solution

After reducing the noise level as described in the data processing section above, I calculated a correlation matrix that contains all the correlations of the fluorescence activity levels among all the neurons. Afterwards, if the correlations met a certain threshold, I predicted that there is a direct connection between the two neurons. The threshold was determined after examining the dataset provided by kaggle that reveals whether two neurons are actually present or not. I used a training set fluorescence_iNet1_Size100_CC01inh.txt, which contained 100 neurons with 179,497 fluorescence fluorescent calcium indicators measured at 20 millisecond time intervals. My test set, fluorescence_iNet1_Size100_CC02inh.txt, was a different neuron network that also contained 100 neurons with 179,497 fluorescence fluorescent calcium indicators measured at 20 millisecond time intervals. After that, I compared the predictions to the actual presence of neuron connections from the file network_iNet1_Size100_CC02inh.txt. The following is an evaluation table, sensitivity value, and 1 – specificity value obtained from the analysis:

Table 1: This is the table of classifications based on predictions

		Classification		
		1	0	
Truth	1	TP = 362	FN = 644	P = 1006
	0	FP = 1252	TN = 17947442	N = 17948694

Table 2: This table contains sensitivity, 1 - specificity, and precision for the predictions made

Sensitivity	0.36
1 – Specificity (False Positive Rate)	6.97*10^(-5)
Precision	0.22

My analysis reveals that the method I used is not nearly as accurate as the winning methods. This can be improved by coming up with a better way of processing the data and by using a different classification method.

My Own Question

How do I incorporate the data including location of neurons to predict the presence of a direct connection? Also, how do I predict the direction of the neuronal connection?

Proposed Solution to Own Question

I do not yet have the skills to do these analyses on my own. However, the winning methods have explored these questions. One method used Random Forest with topological information as a way of incorporating both location of neurons and fluorescence activity to help predict direct connections. Other teams explored the causal relationships of fluorescence activities between two neurons to help determine the directionality of the connections.