EN.580.694: Statistical Connectomics
Final Project Proposal

Jiarui Wang · April 1, 2015

A Statistical Method for Connectome Wide Association Study (CWAS) Validation

Opportunity As CWAS grows as a field, there will be a need for a statistical framework for validating prognosis/diagnosis classification. As more data is generated at higher fidelity, there will be an increasing interest in trying to predict phenotypes based on connectomes. Many classifiers will be created, but there is currently no good way to validate whether or not these classifiers will perform in a real-world setting. A statistical framework will be useful in this kind of validation.

Challenge Current statistical methods rely on parametric tests and asymptotic assumptions, and non-parametric permutation testing have had limited application in the current literature. A main challenge is that there does not exist sufficient data for proper application of these statistical methods in terms of sample quantity and quality.

Action The proposed method aims to bypass these challenges by assuming a null data generating model and permuting null datasets in which the connectomes and phenotypes are actually independent. Both data quantity and quality can be fully controlled under this framework. Classifiers will be tested on the permuted null datasets and optimized to maximize predictive power while minimizing error.

Resolution We will obtain a sandbox statistical environment for which we may control all the hidden parameters of the system for testing classifiers built from real data and to build new classifiers that are valid under the null hypothesis.

Future Work This method can be applied to existing classifiers from the literature in order to assess their correctness. The false discovery rate for these classifiers under the null hypothesis can then be calculated to quantify the mis-classification error.

Statistical Decision Theoretic

A set of adjacency matrices $\mathcal{A}$ and phenotypes y will be sampled randomly and independently according to a SBM model with a block size of 1. Then the classifier of class F will generate estimates of the phenotype y of class $\hat{Z}$ and its performance will be evaluated by the loss function ℓ. This evaluation criteria will be used to assess the correctness of various classifiers of class F.

Sample Space $A \in \mathcal{A}_N$, where Adjacency Matrices: $A = \{0,1\}^{N \times N}$
and $y \in \mathcal{Y}$, where Phenotypes: $y \in \mathcal{R}$ (in this case, the phenotype is continuous, but it is easy to consider a discrete phenotype as a specific case of this sample space)

Model $\{SBM_N^1(\rho, \beta) : \rho \in \Delta_1 \ \beta \in (0,1)\}$

Action Space $\mathcal{Z} \in \mathcal{R}$, where $\mathcal{Z}$ is linked to a phenotype.

Decision Rule Class $F : \mathcal{A}_N \to \mathcal{Z}$

Loss Function $\ell(Z, \hat{Z}) = (Z - \hat{Z})^2$

Risk Function $R = E[\ell(Z, \hat{Z})]$