Declaración del Problema de Tesis

Maximiliano Martino

February 17, 2025

Planteo del Problema

Nuestro objetivo es determinar si dos o más redes cerebrales, representadas como grafos, pertenecen a la misma población o no, con el objetivo adicional de identificar las conexiones específicas (nodos o aristas) responsables de cualquier diferencia detectada. Cada cerebro se modela como un grafo no dirigido $G = (V, E)$, donde V representa regiones fijas del cerebro (nodos), y E representa las conexiones (aristas) entre ellas, que pueden variar entre individuos.

Sea $\mathcal{G}_A = \{G_1^A, G_2^A, \ldots, G_{N_A}^A\}$ una colección de grafos correspondientes a una población de control (cerebros normales), $\mathcal{G}_B = \{G_1^B, G_2^B, \ldots, G_{N_B}^B\}$ una población de cerebros con una enfermedad como la esquizofrenia, y $\mathcal{G}_C = \{G_1^C, G_2^C, \ldots, G_{N_C}^C\}$ una población de cerebros que padecen Alzheimer. Las tres poblaciones tienen el mismo número de nodos $|V|$, pero sus aristas E difieren debido a variaciones en la conectividad cerebral.

Para cada población, definimos un grafo central o representativo, denotado como M_A para la población de control, M_B para el grupo de esquizofrenia y M_C para el grupo de Alzheimer. El grafo representativo M_A se puede obtener promediando las matrices de adyacencia de todos los grafos en $\mathcal{G}_A$, y de manera similar para M_B y M_C.

Para evaluar si un grafo dado G_i^A pertenece a su población respectiva, definimos la distancia $d(G_i^A, M_A)$ como una medida de cuán lejos está este grafo del grafo central de su grupo. De manera similar, podemos calcular la distancia entre cada grafo de una población y el grafo central de otra población, es decir, $d(G_i^A, M_B)$, $d(G_i^A, M_C)$, $d(G_i^B, M_A)$, etc.

Test de Hipótesis

La prueba de hipótesis, desarrollada previamente por otros investigadores, evalúa si las distribuciones de las distancias $d(G, M_A)$, $d(G, M_B)$ y $d(G, M_C)$ son significativamente diferentes entre las poblaciones, permitiéndonos inferir si los grafos provienen o no de la misma población. Esta prueba se construye sobre las siguientes hipótesis:

- **Hipótesis Nula** H_0: Las redes cerebrales de todas las poblaciones (control, esquizofrenia y Alzheimer) provienen de la misma distribución, lo

que significa que no hay diferencias significativas entre los tres grupos. Matemáticamente, esto se expresa como:

$$H_0 : \mathcal{M}_1 = \mathcal{M}_2 = \cdots = \mathcal{M}_m$$

- **Hipótesis Alternativa** H_A: Al menos una de las poblaciones cerebrales es diferente de las demás:

$$H_A : \text{Al menos una } \mathcal{M}_i \neq \mathcal{M}_j$$

La estadística propuesta para probar la hipótesis nula está dada por la fórmula:

$$T = \frac{\sqrt{m}}{a} \sum_{i=1}^{m} \sqrt{n_i} \left(\frac{n_i}{n_i - 1} \overline{d_{G_i}(M_i)} - \frac{n}{n-1} \overline{d_G(M_i)} \right)$$

Donde las variables involucradas son:

- m: El número de poblaciones (por ejemplo, control, esquizofrenia y Alzheimer).
- n_i: El número de grafos (cerebros) en la población i.
- n: El número total de grafos entre todas las poblaciones.
- $\overline{d_{G_i}(M_i)}$: La distancia promedio de los grafos en la población i a su grafo central correspondiente M_i.
- $\overline{d_G(M_i)}$: La distancia promedio total de todos los grafos a el grafo central de la población correspondiente M_i.
- a: Una constante de normalización elegida en función de la forma particular de la estadística de la prueba.

La estadística de prueba T mide la desviación general de las distancias $d(G_i, M_i)$ con respecto a sus valores esperados bajo la hipótesis nula. Donde un T altamente negativo sugiere que los grafos difieren significativamente, con lo cual rechazamos la hipótesis nula a favor de la hipótesis alternativa, lo que significa que las redes cerebrales de las poblaciones comparadas no provienen de la misma distribución.

Objetivo

Sobre la base de esta prueba de detección, nuestro trabajo se centra en *identificar los nodos o conexiones específicos* que son responsables de las diferencias entre las redes de las poblaciones. Nuestra hipótesis es que existen nodos críticos cuya eliminación o modificación reduciría la distancia observada $d(G_i, M)$, acercando más las redes entre sí.

Nuestro enfoque implica dos algoritmos posibles:

1. **Eliminación de Nodos y Re-prueba:** Eliminar nodos iterativamente y recalcular las distancias para identificar nodos críticos que contribuyan a la diferencia.

2. **Eliminación de Aristas y Re-prueba:** Eliminar aristas entre nodos para encontrar diferencias estructurales significativas.

Estos enfoques deberían funcionar bien cuando hay grandes diferencias entre los grupos. Es importante explorar bajo qué leyes de probabilidad o tamaños de grafo este método podría no ser efectivo. Por ejemplo, habría que investigar que pasa cuando hay dependencias entre cierto grupos de aristas ("si aparece la arista k, entonces siempre aparece la arista l")

Desarrollaremos un algoritmo, implementado en `R`, para identificar de manera eficiente los nodos y/o aristas críticas responsables de las diferencias entre las redes cerebrales a nivel poblacional.

Condiciones Exploratorias

1. **Aristas que no aportan información:** Dado que podemos llegar a tener una gran cantidad de aristas, podemos considerar el hecho de eliminar aristas que no aparecen nunca. Es decir, nos quedamos con las aristas positivas.

2. **Separabilidad de T**: Veremos si podemos re-escribir el estadistico de manera recursiva. Si T es separable, es decir, si T de 3 aristas se puede descomponer en T de 2 de esas aristas más una contribución de la tercera, esto podría reducir significativamente la complejidad computacional. Se requiere más análisis para confirmar si esto es aplicable.

Enfoque de Adición de Aristas

Idea: Supongamos que solo tenemos dos grupos, para verificar si son distintos algo natural es basarse en la cantidad de veces que aparece la arista entre el nodo k y el nodo l en los grafos pertenecientes al grupo A, para luego compararlo con las apariciones en los grafos del grupo B:

1. **Cálculo de p-valores:** Para cada arista se calcula un p-valor, utilizando un test ANOVA para proporciones. El enfoque se centra en agregar aristas en lugar de eliminarlas.

2. **Adición de Aristas Basada en p-valores:** El enfoque comienza calculando T considerando solo la arista con el menor p-valor. Esta arista se agrega al conjunto de aristas que se desea descubrir. Luego se agrega la siguiente arista con el segundo p-valor más pequeño, y así sucesivamente, hasta que el incremento en T sea mínimo. Se investigara como definir el criterio de cambio de pendiente.

Desafíos con Diferencias Sutiles

- **Falsos Positivos:** Cuando las diferencias entre los grupos son sutiles, grupos grandes de aristas pueden introducir ruido. Algunas aristas pueden mostrar p-valores pequeños por azar, y estas podrían ser seleccionadas incorrectamente por el algoritmo. Este problema debe ser abordado para asegurar la robustez en la detección de aristas críticas.

Input y Output

Input: Matrices de adyacencia de cada grafo y sus etiquetas de grupo, o un array para cada grupo que contiene las matrices.
Output: Una matriz de adyacencia con 1 en las aristas identificadas por el algoritmo como responsables de las diferencias entre las poblaciones.