**Clusters**

El agrupamiento de datos no es una actividad nueva de nuestra época, sino una actividad que se ha ejercido en la humanidad desde las primeras civilizaciones(o incluso desde antes). Gracias a ella se ha podido describir de una manera más precisa y certera los fenómenos del mundo que nos rodea.

Nuestro entendimiento acerca de cómo nuestro mundo funciona es aún muy limitado, pero hemos podido adquirir conocimientos de problemas complejos, dividiendo esos problemas grandes en problemas pequeños y luego uniendo las partes para así resolverlos. Los clusters en la ciencia de datos se usan para clasificar y analizar de mejor manera la cantidad exponente de datos que sería imposible analizar de una sola tajada.

Para la creación de clusters es necesario tener cierto “background” de los datos que se van a clusterizar, para así poder obtener al final clusters donde sus unidades sean lo más homogéneas posibles dentro de ese cluster y al mismo tiempo sean heterogéneas(diferentes) respecto a otros clusters.

Este resultado ideal no siempre sucede debido a diversas razones: una de ellas es la asignación de atributos/características poco específicas a cada cluster individuales o demasiado generales que al final resulta difícil poder separarlos entre sí. Otra razón puede ser la elección de un método que no es útil en un determinado problema y termina arrojando resultados poco viables. También existen datos que son sumamente difíciles de diferenciar o agrupar por su naturaleza intrínseca. Por último, pueden existir datos que no se puedan separar en grupos.

Los métodos de clustering siempre te arrojan resultados(agrupamientos) incluso aunque no existan grupos diferenciables en los datos, por eso es necesario ser cauteloso al revisar los resultados y no creer que el gato tiene 6 patas cuando en realidad tiene 4. Además, es importante mencionar que diferentes métodos de clustering, incluso usando el mismo set de datos, podrían arrojar diferentes agrupamientos.

No existe un método perfecto para poder clusterizar cada set de datos, puesto que cada set presenta sus propias características y dificultades, pero un buen entendimiento sobre de dónde y cómo se obtuvieron los datos; el conocimiento y la aplicación de diferentes métodos para analizarlos, y posteriormente la comparación de todos los resultados es manera muy conveniente para poder obtener un agrupamiento más cercano a la realidad.

**Árboles**

Obtuvimos 4 diferentes árboles mediante 4 distintos métodos: single, average, complete y ward.D2. No obstante, 3 de los 4 árboles arrojaron resultados similares en la topología principal. Los grupos o clusters grandes que se formaron fueron 4:ABC1, ABC2a, ABC2b y ABC3. En los 3 árboles más similares, ABC1 se encontraba relativamente más cercano a ABC2b, mientras que ABC3 lo era con ABC2a.

El coeficiente aglomerativo, que va de 0 a 1, nos ayuda a medir la cantidad de estructura de agrupamiento encontrada. Un coeficiente con valores cercanos a 1 sugieren una estructura de agrupación fuerte, mientras que valores cercanos a 0 sugieren lo contrario(una agrupación pobre).

Los 3 árboles(average, complete y ward.D2) eran bastantes certeros en cuanto a la topología correcta de las proteínas, pero el más informativo, claro y con el coeficiente aglomerativo más alto(0.91) era el que se obtuvo con el método ward.D2. Respecto a los coeficientes aglomerativos por los métodos average y complete, resultaron ser muy similares(0.63 y 0.70 respectivamente).

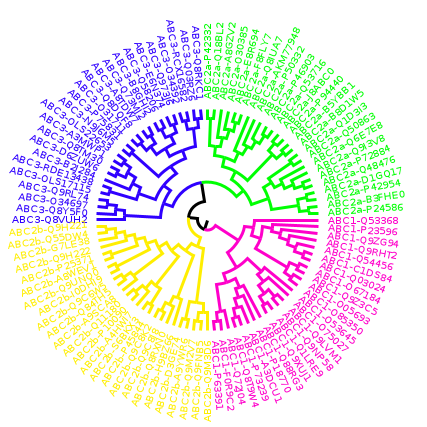
El último árbol,el del método single, fue el menos informativo y el único con la topología diferente a los otros 3. En este árbol, ABC1 y ABC3 eran más cercanos entre sí; ABC2a se subdividió en 2 grupos y un grupo de esa subdivisión formaba un cluster con ABC2b.

La aplicación y comparación de distintos métodos para un set de datos es fundamental para la obtención de una clusterización más cercana a la realidad. Después de hacer la comparación de los métodos anteriores, nos dimos cuenta que 3 de los 4 métodos acertaron con la topología principal de las proteínas correcta, siendo el método de ward.D2 el más preciso y con el coeficiente más alto.

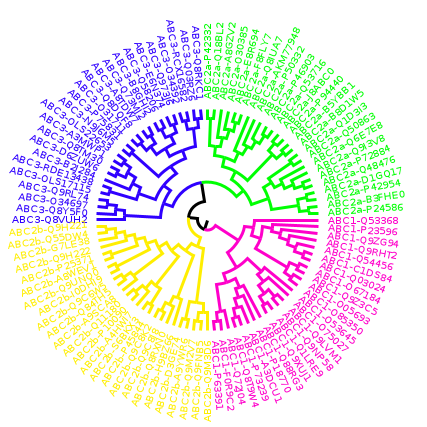
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Método | single | average | complete | ward.D2 |
| Agglomerative coefficient | 0.5298089 | 0.6362812 | 0.7085756 | 0.9180941 |

**Imágenes**

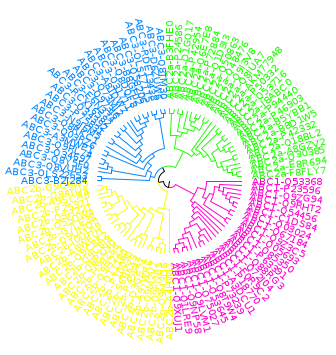
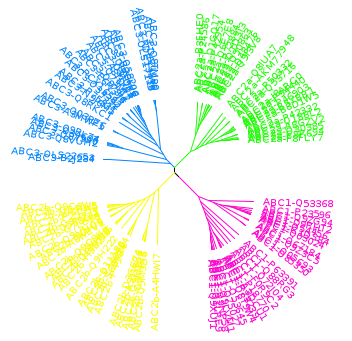
Método -> ward.D2



Método -> complete



Método -> average



Método -> single

