Trabajo Final

Clustering Jerárquico y Medidas de Similitud

Luis David Bassa Llorente, Santiago Enamorado Espitia, Camilo Díaz Torres

Universidad de Córdoba

26 de noviembre de 2024

Contenido

Introducción

En el mundo actual, el análisis de datos es esencial para tomar decisiones informadas en diversas áreas como el marketing, la biología, las ciencias sociales y la salud. Una de las herramientas clave para explorar relaciones y patrones dentro de los datos es el análisis de clúster , que busca agrupar observaciones o variables en subconjuntos homogéneos basándose en sus características compartidas. Para lograr este objetivo, es crucial contar con medidas de similitud , que cuantifican el grado de similitud o diferencia entre los elementos. Estas medidas son el fundamento sobre el cual se construyen los algoritmos de agrupamiento. A partir de ellas, los métodos de clúster jerárquico ofrecen una forma visual y flexible de agrupar los datos, permitiendo identificar relaciones jerárquicas entre los grupos.

Medidas de Similitud

Definición:
Una medida de similitud es una función encargada de evaluar el grado de semejanza entre dos elementos del conjunto de datos que se está analizando. Estos 'elementos' pueden respresentar personas, productos, clientes, genes, o cualquier otra entidad que contenga caracteristicas que puedan ser medidas y comparadas. Este tipo de medida se utiliza para cuantificar qué tan cercanos o similares son dos elementos en función de sus atributos, permitiendonos agrupar estos elementos que puedan ser similares en el respectivo análisis. Las medidas de similitud es bastante utilizada en métodos como el clustering, ya que nos ayuda a agrupar elementos que comparten propiedades comunes.

Fórmulas Principales

- Distancia Euclidiana:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Distancia de Manhattan:

$$d(X, Y) = \sum_{i=1}^{n}|x_i - y_i|$$

- Distancia de Chebyshev:

$$d(X, Y) = \text{máx}(|x_i - y_i|)$$

Fórmulas Principales

- Distancia de Minkowski:

$$d(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- Coeficiente de Coseno:

$$\cos(X, Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Distancia de Jaccard:

$$J(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

Ejemplo en R

```r
##Ejemplo de m didas de similitud
data(iris)
head(iris)
#Seleccionamos solo las primeras cuatro columnas num ricas
datos = iris[, 1:4];datos
#Calulamos la distancia euclidiana entre las observaciones
distancia_euclidiana = dist(datos, method = "euclidean");distancia_euclidiana
distancia_euclidiana = as.matrix(distancia_euclidiana)[1:5, 1:5];distancia_euclidiana
```

	1	2	3	4	5
1	0.0000000	0.5385165	0.509902	0.6480741	0.1414214
2	0.5385165	0.0000000	0.300000	0.3316625	0.6082763
3	0.5099020	0.3000000	0.000000	0.2449490	0.5099020
4	0.6480741	0.3316625	0.244949	0.0000000	0.6480741
5	0.1414214	0.6082763	0.509902	0.6480741	0.0000000

```r
#Calculamos la distancia de Manhattan entre las observaciones
distancia_manhattan = dist(datos, method = "manhattan");distancia_manhattan
distancia_manhattan= as.matrix(distancia_manhattan)[1:5, 1:5];distancia_manhattan
```

	1	2	3	4	5
1	0.0	0.7	0.8	1.0	0.2
2	0.7	0.0	0.5	0.5	0.7
3	0.8	0.5	0.0	0.4	0.8
4	1.0	0.5	0.4	0.0	1.0
5	0.2	0.7	0.8	1.0	0.0

Figura 1: Ejemplo en r medidas de similitud

Interpretacion

La interpretación depende del tipo de método usado, en este caso, la distancia Euclidiana y Manhattan miden la disimilitud; es decir, un valor entre dos observaciones cercano a 0 indican alta similaridad en estas, sus puntos son cercanos. Y un valor cercano a 1 indica que sus puntos están más alejados, es decir, no hay mucha similitud. Contrario a métodos como el coseno, que mide la similitud, quiere decir que un valor cercano a 1 indica que hay bastante similitud, y un valor cercano a 0 indica que no hay mucha similitud. Por otro lado, si tuvieramos que hacer una investigación sobre los datos usados, podríamos utilizar los diferentes métodos mencionados anteriormente con la finalidad de proporcionar información sobre la similitud estructural entre las especies de iris según sus caracteristicas, lo cual podría ser útil en estudios de de botánica, ecología y taxonomía, y sobretodo en proyectos de conservación de biodiversidad o en el desarrollo de herramientas de clasificación automatica

Clustering

Definición:
El clustering es una técnica de aprendizaje no supervisado que agrupa un conjunto de datos en clústeres.o conglomerados, utilizando métodos como las medidas de similitud, donde los elementos dentro de un mismo clúster son más similares entre sí que con los elementos de otros clústeres. El objetivo principal del clustering es descubrir estructuras subyacentes o patrones en los datos sin requerir etiquetas o categorías previas. Existen distintos métodos de clustering, como el clustering jerárquico y el clustering k-means, la elección de estos dependerá del tipo de datos y análisis que se quiera hacer. En esta ocasión hablaremos del clustering jerárquico

Clustering jerárquico

Definición:
El clustering jerárquico es un método de análisis de agrupamiento que busca organizar datos en una jerarquía de grupos o clústeres"según su similitud. Este proceso se representa comúnmente mediante un dendrograma, un gráfico en forma de árbol, donde cada nivel de la jerarquía muestra cómo se fusionan los elementos o grupos en función de su similaridad. Existen varios métodos de enlace, tales como

Definición y Métodos de Enlace

- Agrupa datos en clústeres según su similitud.
- Representado mediante dendrogramas.
- Métodos de enlace:
  - Enlace Completo: Maximiza la distancia entre puntos más lejanos.
  - Enlace Promedio: Usa la distancia promedio entre puntos.
  - Enlace Simple: Minimiza la distancia entre puntos más cercanos.

Dendrograma de Clustering Jerárquico

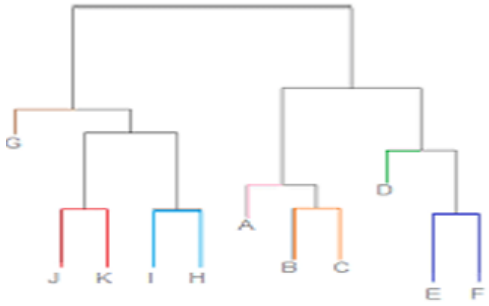

Figura 2: Dendrograma generado en R utilizando enlace completo.

Interpretacion

En un dendrograma, cada elemento en la parte inferior representa una observación individual (en este caso, las letras como J, K, I, H, etc.). A medida que subimos en el árbol, las observaciones comienzan a agruparse en función de su disimilitud. Por ejemplo, las observaciones E y F se fusionan primero, lo que indica que son muy similares. La altura de las fusiones refleja la disimilitud;es decir, las uniones más bajas nos indican una similitud más alta, mientras que uniones más altas (como la de G con el resto) muestran menor similitud. Trazar una línea horizontal a una cierta altura nos permite definir grupos o clústeres específicos

Ejemplo

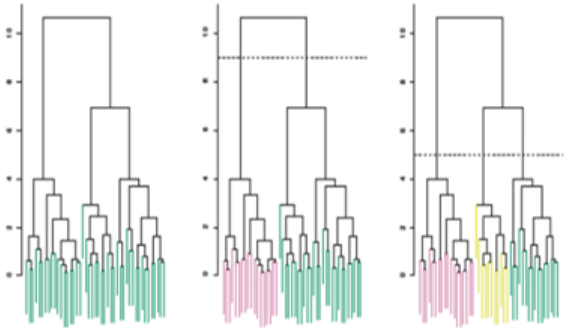

Figura 3: Cortes en dendrograma

Ejemplo de Clustering jerárquico

Para el ejemplo estaremos utilizando la base de datos NCI60 que nos proporciona R. NCI60 es un conjunto de datos que contiene información genética sobre 60 líneas celulares de cáncer humano, recopiladas por el Instituto Nacional del Cáncer de Estados Unidos(NCI). Estas líneas celulares representan distintos tipos de cáncer, como: Leucemia, melanoma, cáncer de mama, cáncer de pulmón, etc.

Enlace Completo

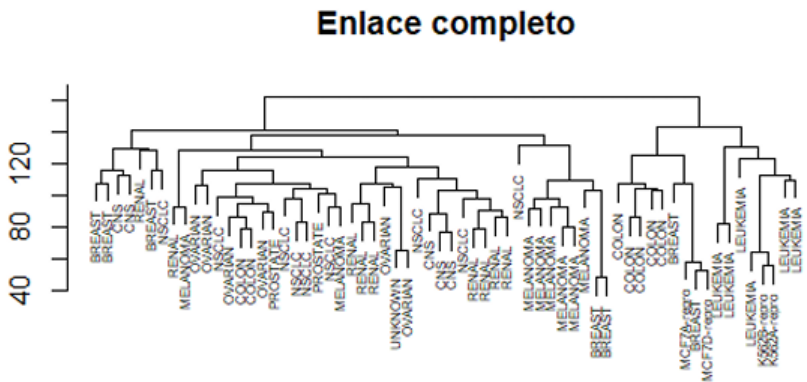

Figura 4: Enlace completo

Enlace promedio

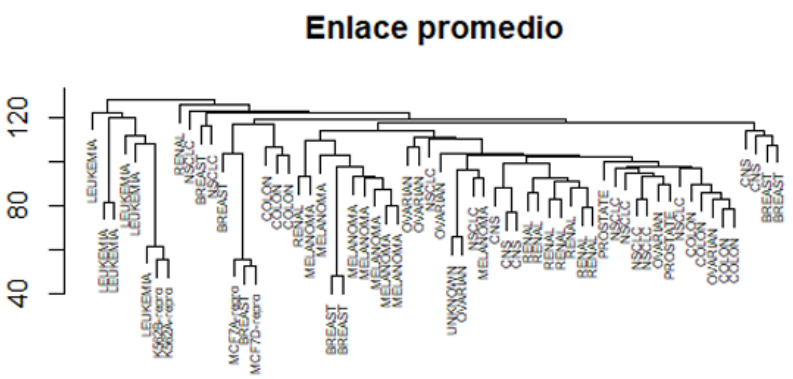

Figura 5: Enlace promedio

Enlace simple

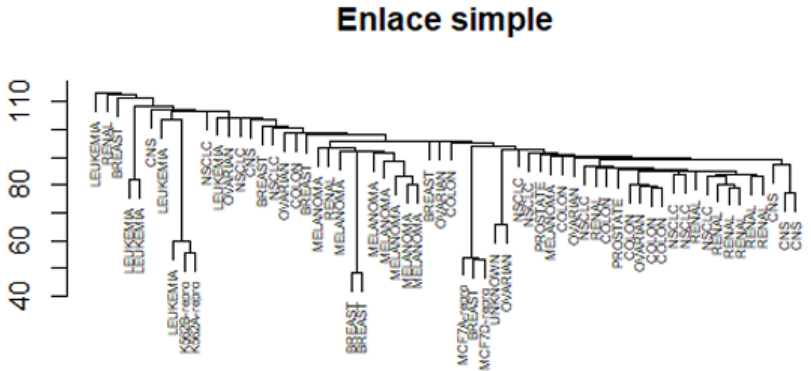

Figura 6: Enlace simple

cluster

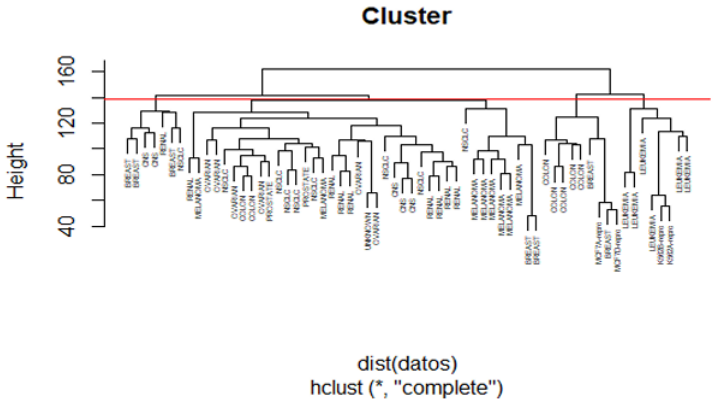

Figura 7: cluster

Interpretación de resultados - Métodos de enlace (Parte 1)

Enlace completo:

- Observando el gráfico 3, al aplicar el enlace completo, se observa una mayor separación entre algunos tipos de cáncer.
- Los grupos generados tienden a tener una alta cohesión interna.
- Tipos de cáncer como *Leukemia* (leucemia) aparecen más separados, lo que sugiere una fuerte disimilitud en la expresión génica.

Enlace promedio:

- Observando el gráfico 4, los grupos tienden a ser menos compactos en comparación con el enlace completo.
- Ciertos tipos de cáncer, como *BREAST* (cáncer de mama) y *COLON* (cáncer de colon), aparecen agrupados con mayor frecuencia.
- Esto podría sugerir similitudes relativas en la expresión génica entre estos tipos de cáncer.

Interpretación de resultados - Métodos de enlace (Parte 2)

Enlace simple:

- Observando el gráfico 5, este método resulta menos consistente en comparación con los otros.
- Agrupa tipos de cáncer como *Leukemia* y *MELAN* (melanoma) de forma más difusa y dispersa.

Consideraciones generales:

- La elección del método de enlace afecta la estructura de los clústeres:
  - *Enlace completo:* Ideal para identificar grupos compactos y bien diferenciados.
  - *Enlace promedio:* Ideal para subgrupos uniformes en similitud.
  - *Enlace simple:* Útil en exploraciones preliminares.
- Observando el gráfico 6, al trazar la línea de corte, se identifican 4 clústeres principales que comparten patrones específicos de expresión génica.

Preguntas sobre medidas de similitud

¿Por qué son importantes las medidas de similitud en el análisis de datos?

- Permiten evaluar el grado de parecido entre distintos objetos o datos en un conjunto.
- Son esenciales en técnicas como el clustering para agrupar objetos similares.
- Ayudan a identificar relaciones y crear grupos homogéneos, clave para entender estructuras y hacer predicciones precisas.

¿Cuál es la diferencia entre la medida de similitud coseno y la distancia Euclidiana?

- *Similitud coseno:* Evalúa el ángulo entre dos vectores, ignorando la magnitud, ideal para datos donde importa más la dirección que la escala.
- *Distancia Euclidiana:* Mide la longitud de la línea recta entre dos puntos, considerando dirección y magnitud. Es útil cuando las diferencias en todas las dimensiones son importantes.

Métodos de enlace y clustering jerárquico

¿Cuáles son las aplicaciones de los métodos de enlace en el clustering jerárquico en el ámbito social?

- *Análisis de redes sociales:* Identificar comunidades basadas en interacciones.
- *Estudios de comportamiento y opinión pública:* Agrupar encuestas o respuestas para identificar patrones.
- *Movilidad y segregación urbana:* Formar clusters de zonas con características similares para diseñar intervenciones.

¿Qué ventajas y desventajas tiene el clustering jerárquico en comparación con otros métodos de clustering, como el K-means?

- *Ventajas:*
  - No requiere especificar el número de clusters de antemano.
  - Maneja estructuras complejas y facilita la visualización mediante dendogramas.
- *Desventajas:*
  - Mayor complejidad computacional.
  - Sensibilidad a valores atípicos.

Preguntas problemáticas

Diferencias entre discriminación y clasificación:

- *Discriminación:* Se enfoca en identificar características distintivas entre grupos conocidos.
- *Clasificación:* Asigna nuevos elementos a categorías basándose en patrones aprendidos.

Determinar el número óptimo de grupos en clustering:

- *Método del codo:* Identifica el número óptimo donde la reducción de la suma de distancias comienza a ser mínima.
- Otras técnicas: Validación cruzada y análisis de variabilidad entre y dentro de los clusters.

Desafíos éticos en técnicas de discriminación y clasificación:

- Posible refuerzo de estereotipos y desigualdades debido a sesgos en los datos.
- Preocupaciones sobre privacidad y consentimiento informado.
- Soluciones: Auditorías de algoritmos, transparencia y representatividad en los datos.

Bibliografía

- Cristina Gil. (s.f.). *Clustering.* Publicado en RPubs.
- Irani, K. (2016). *Clustering Techniques in Data Mining.* International Journal of Computer Applications.