Cluster Analysis

Daniel Czarnievicz

Descripción general

El objetivo general del análisis de clusters es formar grupos (**clusters**) de acuerdo a la algunas características de interés. El procedimiento general puede describirse de la siguiente forma:

- Se parte de la matriz de datos $\boldsymbol{X}_{I\times J}$, generalmente estandarizada.
- Se define alguna forma de medir similitudes entre los individuos (medida de distancia).
- Se crea la matriz $\boldsymbol{D}_{I\times I}$ donde el elemento d_{ij} mide el grado de similitud entre los individous i y j.
- Se definen algoritmos de clasificación.
- Se definen stopping rules.
- Se selecciona el número de grupos en función de las características consideradas.

El análisis de clusters forma parte de los llamados métodos de clasificación no supervisada.

La metedología también puede aplicarse por variables, en lugar de por observaciones. El objetivo en estos casos es detectar similitudes o jerarquías entre las variables para luego utilizar métodos de reducción de dimensionalidad.

Medidas de distancia

La similitud entre objetos es una medida de correspondencia, asociación o parecido entre objetos que van a ser agrupados. La semejanza puede ser definida mediante una función $s_{ij} : \mathbb{R}^J \to \mathbb{R}$, donde $s_{ij} = s_{ji} \ \forall i \forall j$. Para que s_{ij} sea un índice de similitud debe cumplirse que $s_{ij} \leq s_{ii} = s_{jj} \ \forall i \forall j$, y, en general, $0 \leq s_{ij} \leq 1$.

Si s_{ij} es un índice de similaridad, entonces $d_{ij} = 1 - s_{ij}$ es un índice de disimilaridad con:

- $0 \leq d_{ij} \leq 1$

- $d_{ij} = 0 \Leftrightarrow i = j$, por lo tanto, $d_{ii} = d_{jj} = 0 \ \ \forall i \forall j$
- $d_{ij} = d_{ji}$ (propiedad de simetría).

Llamaremos **distancia (en el espacio métrico E)** a las disimilariadades que satisfacen:

i) $d_{ij} > 0 \ \ \forall (i,j) \in E \ \text{ con } \ i \neq j$

ii) $d_{ij} = 0 \Leftrightarrow i = j \forall (i,j) \in E$

iii) $d_{ij} \leq d_{ik} + d_{kj} \forall (i,k,j) \in E$ (desigualdad triangular)

Algunas medidas de distancia comúnmente utilizadas

Para variables cuantitativas y siendo p el número de variables:

$$\boxed{\text{Dist. Euclídea: } d_{ij}^2 = \sum_{k=1}^{p} \Big(x_{ik} - x_{jk}\Big)^2}$$

$$\boxed{\text{Dist. Euclídea Reducida: } d_{ij}^2 = \sum_{k=1}^{p} \frac{\Big(x_{ik} - x_{jk}\Big)^2}{s_k^2}}$$

$$\boxed{\text{Dist. Minkowski de orden t: } d_{ij}^2 = \sum_{k=1}^{p} \frac{\Big|x_{ik} - x_{jk}\Big|^t}{s_k^t}}$$

$$\boxed{\text{Dist. Mahalanobis: } d_{ij}^2 = \Big(x_{ik} - x_{jk}\Big)' \mathbf{\Sigma}^{-1} \Big(x_{ik} - x_{jk}\Big)}$$

Variables cualitativas

Cuando en los datos se consideran varias variables cualitativas y cuantitativas, las distancias antes descriptas no son buenas medidas de la disimilaridad entre los elementos muestrales. Para poder trabajar con ellas es construye el índice de similaridad global, a partir de los índices por variables. La similariadad entre los elementos i y h en las $j = 1; \ldots, p$ variables se define como:

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih}\, s_{jih}}{\sum_{j=1}^{p} w_{jih}}$$

donde s_{jih} es el índice de similaridad entre los elementos i y h en la variable j, y w_{jih} es el peso asignado a la variable j, pudiendo este ser incluso 1 o 0.

Los índices de similariadad para cada tipo de variable se construyen de la siguiente forma:

- **Variables cualitativas**: puede construirse por bloque o para cada variable.
  - Cuando se realiza por variable la similitud será 1 en los casos en que ambas unidades posean o no el atributo, y 0 en caso de que una de ellas lo posea y la otra no.
  - Cuando se las trata de forma conjunta se construyen tablas de asociación contando los atributos presentes en:
    - ambos elementos: a
    - en i y no en h: b
    - en h y no en i: c
    - en ningún elemento: d

    Luego la similitud puede construirse mediante:
    - *Prop. de coincidencias*: se calcula omo el número total de coincidencias sobre el número total de atributos.
    $$s_{ij} = \frac{a+d}{n_a}$$
    - *Prop. de apariciones*: proporción de veces donde el atributo aparece en ambas observaciones.
    $$s_{ij} = \frac{a}{a+b+c}$$
- **Variables cuantiativas**:
$$s_{jih} = 1 - \frac{|x_{ij} - x_{hj}|}{rg(x_j)}$$

Una vez construidas las similaridades la distancia puede defirnirse como $d_{ij} = 1 - s_{ij}$, pero esta puede no cumplir la propiedad triangular. Para los casos en que la matriz sea semi definida positiva, dicha propiedad sí se cumplirá si se calcula la distancia como:

$$d_{ij} = \sqrt{2(1 - s_{ij})}$$

Grupos

Existen distintos métodos a través de los cuales constriur los grupos. Los mismos se clasifican según sean *divisivos* o *agregativos*, y según sean *jerarquicos* o *no jerarquicos*.

- **Métodos divisivos**: se parte de un solo grupo con I individuos y se particiona hasta obetener I grupos con 1 individuo cada uno.

- **Métodos agregativos**: se parte de I grupos con 1 individuo cada uno y se agregan hasta obtener 1 grupo con I individuos.

- **Métodos jerárquicos**: genera particiones solapadas. No permite la reasignación de unidades. Estrictamente, no genera grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos. Dicha jerarquía puede utilizarse para conformar grupos.

- **Métodos no jerárquicos**: se predetermina la cantidad de grupos no solapados. Permite la reasignación de unidades.

Métodos jerárquico-agregativos.

En los métodos jerárquico-agregativos se parte de I grupos con un individuo cada uno, los cuales se agregan hasta obtener 1 grupo con I individuos. En cada paso, los individuos o grupos más parecidos se unen para formar un nuevo grupo. Esto implica que los individous agrupados en pasos anteriores no pueden cambiar de grupo (a esto nos referimos cuando hablamos de particiones no solapadas).

El primer paso es siempre igual y muy sencillo. Se parte de la matriz de datos $\boldsymbol{X}_{I\times J}$, y se construye la matriz de distancia $\boldsymbol{D}_{I\times I}$ según la métrica seleccionada. Luego, las dos unidades más parecidas (es decir, la de menor distancia), se unen para formar un grupo.

El problema comienza con los pasos subsiguiente cuando deben unirse observaciones con grupos previamente formados, o grupos con grupos. La pregunta a la que se debe dar respuesta es ¿cuál es la distancia entre la observación k, y el grupo formado por las observaciones i y j? (Análogamente uno podría preguntarse cuál es la distancia entre el grupo U_1 y el grupo U_2, siendo estos dos grupos formados en pasos anteriores del algoritmo). Para definir estas distancias existen distintos criterios:

- **Single Linkage**: $d_{(i,j),k} = \min\{d_{i,k};\, d_{j,k}\}$

- **Complete Linkage**: $d_{(i,j),k} = \max\{d_{i,k};\, d_{j,k}\}$

- **Average Linkage**: la distancia entre dos grupos es el prmedio entras las distancias entre pares de observaciones. Es sesgado hacia la formación de grupos con igual varianza.

- **Centroide**: define la distancia entre los grupos K y L como $d_{(K,L)} = ||\bar{x}_K - \bar{x}_L||^2$. Tiene el beneficio de ser robusto a la presencia de outliers.

- **Ward**: la unión entre grupos se relaiza de forma tal de minimizar la varianza interna.

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-\bar{x}_j)^2}_{\text{T (Total variance)}} = \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{J}\left(x_{ij}^{(k)}-\bar{x}_j^{(k)}\right)^2}_{\text{W (Within group variance)}} + \underbrace{\sum_{k=1}^{K}\sum_{j=1}^{J}n_k\left(\bar{x}_j^{(k)}-\bar{x}_j\right)^2}_{\text{B (Between group variance)}}$$

  Esto puede verse en términos de distancias realizando las sumas en J:

  - T es la suma de las distancias entre cada una de las i observaciones y el centroide de las observaciones, es decir: $\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-\bar{x}_j)^2 = \sum_{i=1}^{I} d^2_{(i,G)}$
  - W es la suma de las distancias entre cada una de las observaciones y el centroide del grupo k, es decir: $\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{J}\left(x_{ij}^{(k)}-\bar{x}_j^{(k)}\right)^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k} d^2_{(i,G_k)}$
  - B es la suma de n_k-veces las distancias entre el centroide del grupo k y el centroide total de la observaciones, es decir: $\sum_{k=1}^{K}\sum_{j=1}^{J} n_k\left(\bar{x}_j^{(k)}-\bar{x}_j\right)^2 = \sum_{k=1}^{K} n_k d^2_{(G_k,G)}$.

  El crecimiento de la inercia intraclase puede medirse mediante:

$$\Delta_{(K\cup L)} = \frac{n_K\, n_L}{n_K + n_L}\, d^2_{(G_K,G_L)}$$

  El algoritmo de Ward es sesgado hacia la formación de grupos de igual tamaño.

Características generales:

- Ward y Centroide tienden a formar grupos más esféricos. Son menos sensibles a la presencia de outliers.
- Complete Linkage tiende a producir grupos esféricos de diámetro muy parecidos.
- Single Linkage es más sensible a la presencia de outliers (los cuales generan un efecto cadena). El algoritmo tiende a separar a los outliers dejando grupos unitarios para el final.

Métodos jerárquico-divisivos.

Trabaja en dirección opuesta a los métodos agregativos. Es decir, parten de 1 grupo con I obersvaciones y en cada paso dividen el grupo jerarquico del que parten buscando construir los grupos más disímiles.

Stopping Rules

Un problema en el análisis de clusters mediante métodos jerárquicos consiste en poder determinar cuántos grupos deben formarse. Es decir, cuándo debe frensarse el algortimo de unión y proceder a describirse los grupos hasta entonces formados. Para ayudar con esto se recuerre a los siguiente indicadores:

- Inspección visual del dendrograma.
- $\boldsymbol{R^2}$: el índice se calcula a cada paso teniendo en cuenta la relación entre la varianza interna (W) y la varianza total (T): $R^2 = 1 - \frac{W}{T}$. Cuando se tienen I grupos de 1 individuo, $R^2 = 1$. Cuando se tiene un grupo de I individuos, $R^2 = 0$. Si al pasar de $k+1$ grupos a k grupos el aporte al R^2 no es significativo, nos quedamos con $k+1$ grupos.
- $\boldsymbol{pseudo - F}$:

$$F_{p(k-1);\,p(n-k)} = \frac{tr(B)/(k-1)}{tr(W)/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

  - Si F crece de forma monótona al crecer la cantidad de grupos k, entonces no se puede determinar una clara estructura de grupos.
  - Si F decrece de forma monótona al crecer la cantidad de grupos k, entonces no se puede determinar una clara estructura de grupos, pero se puede decir que existe una estructura jerárquica.
  - Si F presenta un máximo en k grupos, entonces la población presenta un número definido de grupos en dicho máximo.

- $\boldsymbol{pseudo - t^2}$:

$$t^2 = \frac{tr(W_{G,L}) - [tr(W_G) + tr(W_L)]}{[tr(W_G) + tr(W_L)]/(n_G + n_L - 2)}$$

  Si al pasar de $k+1$ a k grupos el ínidic presenta una caída muy grande, entonces nos quedamos con $k+1$ grupos. La idea es que el aumento de la heterogeneidad es demasiado grande y no conviene unir los grupos.

Métodos no jerárquico.

Los métodos no jerárquicos son también conocidos como métodos de partición, dado que se basan en particionar los datos en una cantidad predeterminada de grupos, G. Una forma de trabajar este problema sería construir todos los posibles G grupos en los que las n observaciones podrían particionarse. El problema de esto es que la cantidad posible de grupos a estudiar es prohibitivamente grande incluso para valores moderados de n.

k-medias

El algoritmo más utilizado para realizar dichas particiones es conocido como **k-medias**, y requiere de las siguientes cuatro etapas:

- Seleccionar G puntos como centros de los grupos.
- Calcular la distancia euclidea de cada observación a los centros de cada grupo, y asignar cada una de ellas al centro más próximo. Cada vez que una observación es asignada a un grupo, se re calcula el centro de dicho grupo.
- Definir un criterio de optimalidad y comprobar si reasignar uno a uno cada elemento de un grupo a otro mejora el criterio.
- Si no es posible mejorar el criterio, terminar el proceso.

El método es sensible a la elección de los centros iniciales. Siempre se recomienda realizar el procedimiento con varios sets de centros iniciales. Si los resultados cambian drasticamente entre un set y otro, o si el algoritmo demora mucho tiempo en converger, puede deberse a que no exista una estructura de grupos en los datos.

k-medias también puede utilizarse en conjunto con métodos jerárquicos de la siguiente forma. Luego de construidos los clusters mediante algún método jerárquico, los centros de dicho grupos se utilizan como puntos iniciales para el algoritmo.

Un criterio de optimalidad u homogeneidad comúnmente utilizado es minimizar la *suma de cuadrados dentro de los grupos* (SCDG)[1]:

$$\min\{SCDG\} = \min\left\{\sum_{g=1}^{G}\sum_{j=1}^{p}\sum_{i=n_g}^{G}(x_{ijg} - \bar{x}_{jg})^2\right\} = \min\left\{\sum_{g=1}^{G}\sum_{j=1}^{p} n_g\, s_{jg}^2\right\}$$

donde x_{ijg} es el valor de la variable j en el elemento i del grupo g, $\bar{x}_{jg}$ es la media de dicha variable en dicho grupo, n_g es la cantidad de observaciones en el grupo g, y s_{jg}^2 es la varianza muestral de la variable j en el grupo g.

Dado que entontrar la optimalidad de este criterio implicaría calcularlo para todas las posibles particiones de las n observaciones, se agrega la restricción de que en cada iteración del algoritmo, solo una observación sea reasignada. El algoritmo se implementa de la siguiente forma entonces:

- Partir de una asignación inicial.
- Comprobar si moviendo algún elemento se reduce $\boldsymbol{W}$.
- Si es posible, mover el elemento, recalcular las medias de los dos grupos afectados por el cambio, y volver al paso anterior. Si no es posible reducir $\boldsymbol{W}$, terminar.

[1] Este criterio equivale a minimizar la traza de la matriz de varianzas internas de los grupos, $\boldsymbol{W}$.

Este mismo algoritmo puede utilizarse tomando otro tipos de centros. Por ejemplo, k-medoides toma como centros iniciales a los individuos representativos de cada grupo. Por representativos nos referimos a los individos que minimizan la disimiliaridad promedio a todos los objetos del cluster. De este forma se logra que los centros, efectivamente sean individuos pertenecientes a la muestra y no puntos ficticios. El método es menos sensible a la presencia de outliers.

Para definir los medoides, supongamos que $x_1, \ldots, x_n$ son puntos en un espacio métrico (X, d). Entonces, el medoide se define como:

$$x_{medoide} = \underset{y \in \{x_1, \ldots, x_n\}}{\arg\min} \sum_{i=1}^{n} d(y, x_i)$$

KNN

Clasificador de Bayes

Metodología

En la práctica, no es posible conocer la distribución de $Y|X$, y por lo tanto, el clasificador de Bayes no puede ser calculado. KNN estima la distrubición de $Y|X$, y asigna las observaciones a la clase con mayor probabilidad estimada.

Dado un $K \in \mathbb{N}$ y una observación x_0, KNN primero identifica los k vecinos más cercanos. Estos constituyen el conjunto $\mathcal{N}_0$. Luego se estima la probabilidad condicional para la clase j como la fracción de puntos en $\mathcal{N}_0$ cuya variable de respuesta es igual a j, esto es:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathrm{I}_{(y_i = j)}$$

Cuando $K = 1$, KNN es demasiado flexible y encuntra patrónes en los datos que no se corresponden con el clasificador de Baye. El error de training es cero, pero el error en el grupo de test puede ser muy alato. A medida que K aumenta, el método se vuelve menos flexible y eventualmente produce un clasificador que es lineal.

K	Varianza	Sesgo
1	muy alta	bajo
crece	baja	alto

Clusters basados en modelos

La clusterización basada en modelos es un método en el cual se asume la existencia de un modelo matemático, y se busca optimizar el ajuste entre el modelo y los datos. Generalmente, dicho modelo es una mezcla de distribuciones. Cada distribución determina la probabilidad de que una observación tenga un conjunto particular de atributo-valores, dado que pertenece a una de las k distribuciones.

Algorítmo EM (Expectation Maximazation)

El algoritmo EM busca la estimación de los parámetros de las distribuciones que mejor se adaptan a los datos y al modelo propuesto (máxima verosimilitud). Con ello se estima el grado de pertenencia de cada observación a cada grupo.

El algoritmo comienza con la estimación inicial de los parámetros de las distribuciones (maximization) y los utiliza para calcular las probabilidad de que cada observación pertenezca a un cluster (expectation). Luego utiliza dichas probabilidades para re-estimar los parámetros de las distribuciones. El procedimiento se repite hasta converger. El ajuste global del modelo se evalúa a través del BIC.

Fuzzy sets

La clusterización fuzzy es una generalización de los métodos de partición. En estos, cada observación es asignada a un y solo un grupo. Los algoritmos fuzzy se basan en computar un *coeficiente de membresía* para cada observación y cada cluster. El coeficiente toma valores entre 0 y 1, de forma tal que la suma de todos los coeficientes para una misma observación, suman 1.

El algoritmo fuzzy no utiliza individuos representativos, sino que busca minimizar la función objetivo:

$$C = \sum_{v=1}^{k} \frac{\sum\limits_{i,j=1}^{n} u_{iv}^2 \, u_{jv}^2 \, d(i,j)}{2 \sum\limits_{j=1}^{n} u_{ju}^2}$$

donde u_{iv} es el coeficiente de membresía de la obserción i en el cluster v. La expresión del denominador aparece multiplicada por dos dado que se están sumando tanto el par (i, j) como el (j, i). A su vez, dado que estamos sumando sobre todos los clusters posibles, la función es entonces una medida de dispersión total entre dichos clusters. El algoritmo itera hasta que el cambio en la función objetivo es menor a una tolerancia ε.

Algunas clusterizaciones son más difuzas que otras. Cuando un individuo tiene igual coeficiente de membresía en todos los clusters (todos iguales a $1/k$), decimos que presenta *complete fuzziness*. Por otro lado, cuando un individuo tiene coeficiente 1 en un cluster y 0 en todos los demás, estamos frente a un caso de partición. Para medir que tan rígida es laa

clusterización se utiliza el coeficiente de partición de Dunn:

$$F_k = \sum_{i=1}^{n} \sum_{v=1}^{k} \frac{u_{iv}^2}{n}$$

Cuando la clusterización es completamente disfuza, $F_k = 1/k$, mientras que cuando la clusterización es completamente particionada, toma valor 1. La versión normalizada del coeficiente de Dunn viene dada por la fórmula:

$$F_k' = \frac{F_k - (1/k)}{1 - (1/k)} = \frac{k\,F_k - 1}{1 - k}$$

la cual siempre toma valores entre 1 y 0, independientemente de la cantidad de clusters k elegidos.

Habitualmente, las observaciones son asignadas al cluster para el cual tienen mayor coeficiente de membresía.

Silhouette

Un gráfico de Silhouette es una forma de validar el resultado de una clusterización. En el mismo se mide la cohesión (que tan similar es cada observación a los miembros de su propio cluster), y la separación (que tan disimil es cada observación a los miembros de los demás clusters). El Silhouette toma valores entre -1 y 1, donde valores altos indican que la observación está bien correspondida con los elementos de su propio cluster, y mal matcheada con los elementos de los demás clusters. Si la mayoría de las observaciones tienen valores altos de Silhouette, entonces la clusterización es considerada buena. Si, en cambio, varias observaciones tienen valores muy bajos, entonces es evidencia de que se consideraron ya sea muchos o muy pocos clusters.

La Silhouette se calcula con una métrica de distancia. Así, dada una clusterización de k clusters, definimos $a(i)$ como la distancia media entre la observación $i \in \mathcal{C}_i$, es decir:

$$a(i) = \frac{1}{|\mathcal{C}_i| - 1} \sum_{j \in \mathcal{C}_i,\, i \neq j} d(i,j)$$

$a(i)$ puede interpretarse entonces como una medida de qué tan bien asignada está la observación i en el cluster $\mathcal{C}_i$.

De forma similar, definimos $b(i)$ como la menor disimilaridad promedio entre la observación i, y los miembros de los demás clusters $\mathcal{C}$ con $\mathcal{C} \neq \mathcal{C}_i$.

$$b(i) = \min_{k \neq i} \left\{ \frac{1}{|\mathcal{C}_k|} \sum_{j \in \mathcal{C}_k} d(i,j) \right\}$$

Definimos la medida de silhouette para la observación i, como:

$$s(i) = \begin{cases} \dfrac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} & \text{si} \quad |\mathcal{C}_i| > 1 \\ 0 & \text{si} \quad \mathcal{C}_i = 1 \end{cases}$$

El promedio de $s(i)$ en todos las observaciones de un cluster es una medida de qué tan compactamente armado está el grupo. El promedio $s(i)$ sobre todos los datos, es una medida de qué tan buena es la clusterización.

Referencias

Beygelzimer, Alina, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. 2018. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. https://CRAN.R-project.org/package=FNN.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. Wiley New York.

Scott, David W. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Wasserman, Larry. 2007. *All of Nonparametric Statistics*. Springer, New York.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.