Clustering

Curso 2019/2020

Sofía Almeida Bruno
Daniel Bolaños Martínez
José María Borrás Serrano
Fernando de la Hoz Moreno
Pedro Manuel Flores Crespo
María Victoria Granados Pozo

Índice

1. Introducción

El clustering consiste en agrupar objetos similares. Dos objetos se consideran similares si, considerando alguna medida de error, las observaciones podrían ser del mismo objeto. Esta clasificación ocurre constantemente en nuestra vida diaria, damos el mismo nombre a objetos que difieren en detalles insignificantes.

Si tenemos una serie de observaciones sin clasificar, el objetivo del clustering es agrupar los datos en clases o clusters. Por ejemplo, cuando en ámbitos biológicos se quiere determinar las especies de una planta concreta. Este tipo de problema aparece cuando queremos no solo identificar especies nuevas, sino también cuando queremos establecer las relaciones entre ellas. El término clustering se considera sinónimo a taxonomía numérica o clasificación. En el ámbito de la ciencia de datos se considera problemas diferentes clasificación y agrupamiento. En el primero conocemos de antemano las clases de los objetos, mientras que en el segundo, a partir de los grupos creados se inferirán las características principales de los grupos.

Utilizando el lenguaje matemático, podemos definir el problema como sigue:

> Dadas $x_1, \ldots, x_n$ medidas de p variables en n objetos considerados *heterogéneos*. El objetivo del análisis cluster es agrupar estos objetos en k clases *homogéneas*, donde k es también desconocido (aunque habitualmente se asume que es mucho menor que n).

Decimos que un grupo es *homogéneo* si sus miembros están cerca unos de otros pero los miembros de otros grupos son muy diferentes a estos. Esto lleva a definir dos métricas entre los puntos para indicar el grado de alejamiento y el de asociación o similaridad. Se pueden tomar distintas distancias, creando aproximaciones diferentes al problema (trataremos este tema en más profundidad en la Sección 2).

El análisis cluster se aplica en numerosos campos como las ciencias naturales, médicas, económicas, *marketing*, ... En *marketing*, por ejemplo, es útil dividir a los clientes y conocer las necesidades de cada segmento de mercado para lograr alcanzar a los clientes potenciales. En psicología puede ser útil encontrar tipos de personalidad a partir de los cuestionarios realizados. En arqueología se puede aplicar esta técnica para clasificar objetos en diferentes periodos.

Para llevar a cabo un análisis cluster hay que realizar principalmente dos pasos:

1. Elegir una medida de proximidad.
2. Elegir un algoritmo para construir los grupos. Hay dos grupos principales: de particionamiento y jerárquicos (que a su vez se dividen entre divisivos y aglomerativos)

2. Medidas

3. Métodos de agrupamiento

3.1. Jerárquicos

3.2. No jerárquicos

4. Número de clústeres

En los algoritmos de *clustering*, uno de los problemas es determinar el número de clústeres k y que es distinto al propio proceso de agrupamiento. Este procedimiento conlleva tener en cuenta varios índices.

Evaluar la solución obtenida para k clústeres es similar a la práctica de determinar la dimensión del espacio en el análisis de componentes principales o el orden de los factores en análisis factorial. La correcta elección de k suele ser ambigua ya que depende de las interpretaciones según la forma y la escala de de la distribución de los datos y la solución deseada. También hay que tener en cuenta que incluso con datos aleatorios se pueden detectar falsos clústeres.

En cada paso de nuestro proceso se crea un nuevo clúster formado por dos observaciones, una observación y otro clúster o dos clústeres ya obtenidos. Así, el número de clústeres k decrece de n (número de observaciones) a 1. La distancia entre dos clústeres es la euclídea o la de disimilitud. Por ejemplo, para los enlaces únicos, completos o de medias las distancias son el mínimo, máximo y la euclídea media respectivamente. Para el método de Ward, es la suma de las distancias entre clústeres al cuadrado dada en la fórmula (REFERENCIAR). Como k decrece de n a 1, el valor de la distancia debería aumentar ya que tendría que ser mayor cuando dos clústeres distintos se agrupan en uno solo.

Uno de los procedimientos para determinar el número de clústeres es el denominado "método del codo" o *elbow method*. Suele ser ambiguo y no muy fiable por lo que se recomienda el uso de otras técnicas. Consiste en dibujar la gráfica de la varianza de la distancia a los centros de cada clúster en función del número de clústeres. En un punto, se observará que no hay demasiada mejora en añadir un clúster más por lo que se escoge dicho punto.

PONER EJEMPLO

Hemos construido una especie de índice de separación o test de modo que si lo dibujamos en función del número de clústeres nos ayuda a identificar visualmente el k adecuado.

En análisis de regresión, el coeficiente de determinación, R^2, es una medida de la varianza total de la variables dependientes según las independientes. En *clustering* debemos construir un índice de R^2 que varíe en base al número de clústeres. Para n clústeres la suma total de las distancias al cuadrado es $T = \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2$ y dentro de de un clúster C_k es $SSE_k = \sum_{i=1} = \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2$. Así, para k clústeres definimos R^2 como

$$R_k^2 = \frac{T - \sum_k SSE_k}{T}.$$

Para n clústeres $SSE_k = 0$ por lo que $R^2 = 1$. A medida que vamos realizando los agrupamientos, estos estarán más separados. Una gran disminución en R_k^2 representaría un agrupamiento diferente. También podríamos tener en cuenta el cambio en R^2 al unir los clústeres R y S como $SR^2 = R_k^2 - R_{k-1}^2$. El estadístico SR^2 representa la proporción de $SSE_t - (SSE_r + SSE_s)$ donde los clústeres C_R y C_S se han unido para formar el clúster C_T.

El objetivo del análisis clúster es encontrar en menor número de clústeres homogéneos. Para un solo clúster, la varianza agrupada para todas las variables es la media de las varianzas de cada una de las variables, es decir, $s^2 = \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 / p(n-1)$. También podemos calcularla para un clúster C_k con n_k observaciones de la siguiente manera:

$$s^2 = \sum_{i=1}^{n_k} \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2 / p(n_k - 1).$$

Grande valores de la varianza agrupada indica que los clústeres no son homogéneos. Por lo tanto, si tiene hacia cero para algún $k < n$ indica la formación de un clúster homogéneo.

Bajo una normal multivariante e independencia de los n vectores de dimensión p para $\Sigma = \sigma^2\mathbf{I}$, podríamos hacer una prueba para verificar que los k clústeres muestran una separación significativa usando, por ejemplo, un estadístico $ANOVA\ F$ (?????). También podemos comprobar si dos medias

están lo suficientemente separadas en cualquier nivel usando un estadístico t (??????). Como la independencia y la distribución normal multivariante no se suelen dar juntos, los estadísticos se denominan pseudo estadísticos F y t^2. El pseudo estadístico F se define como

$$F_k^* = \frac{(T - \sum_k SSE_k)/(k-1)}{\sum_k SSE_k/(n-k)}.$$

Si F_k^* disminuye con k, no deberíamos usarlo para estimar k. Sin embargo, si F_k^* disminuye con k y alcanza máximo, el valor de k en el máximo o el inmediatamente anterior al punto será el candidato del número de clústeres. Por otro lado, el pseudo estadístico t^2 se define como

$$p - t^2 = \frac{[SSE_t - (SSE_r + SSE_s)](n_R + n_S - 2)}{SSE_r + SSE_s},$$

para agrupar los clústeres C_R y C_S con n_R y n_S elementos respectivamente. De nuevo, uno puede construir la gráfica con los valores obtenidos y el número de clústeres. Si los valores son irregulares en cada punto de agrupamiento, no es un buen índice. Pero si la gráfica parece un palo de hockey el valor $k+1$ que causa que la pendiente cambie es nuestro candidato a número de clústeres.

Varios estadísticos se generan por los programas que realizan el proceso de agrupamiento y son dibujados para evaluar heurísticamente el número de clústeres generados. Otras técnicas utilizadas para determinar el k óptimo son por ejemplo el el método de la silueta (*silhouette method*) o el de la brecha (*gap*).

En el *silhouette method*, se observa la similitud de cada observación con su clúster en comparación con el resto de clústeres. Si el índice se encuentra entre los valores -1 y 1 donde un valor próximo a 1 significa un buen agrupamiento. Definimos el índice en este método para cada observación i como

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{b(i), a(i)\}}, \quad \forall i = 1, \ldots, n$$

donde $a(i)$ es la media de la disimilaridad entre i y el resto de puntos que pertenecen al mismo clúster, es decir,

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

y $b(i)$ como la menor distancia media de i a todos los puntos de los clústeres al que i no pertenece:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j).$$

Destacamos que si $|C_i| = 1$ entonces $s(i) = 0$. Así, si muchos objetos tienen un valor alto indica que el resultado obtenido es satisfactorio por lo que se escoge el k que maximice el valor medio de $s(i)$.

Finalmente, el método de la brecha es parecido al método del codo y consiste en comparar la variación total dentro de un clúster para diferentes valores de k con sus valores esperados. El k elegido será aquel que maximice el valor de la brecha. Definimos el estadístico como:

$$Gap(k) = E_n^*\{\log(W_k)\} - \log(W_k).$$

En la fórmula anterior E_n^* denota la media de una de muestra de tamaño n y

$$W_k = \sum_{R=1}^{k} \frac{1}{2n_R} \sum_{ij \in C_R} d(i,j).$$

Notamos que se puede usar para cualquier método y distancia. Según [5] el número 2 en la fórmula de W_k es para que funcione correctamente.

5. Parte práctica

6. Bibliografía

Referencias

[1] Determining the number of clusters in a data set. https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set. Último acceso: 28/12/2019.

[2] Elbow method (clustering). https://en.wikipedia.org/wiki/Elbow_method_(clustering). Último acceso: 28/12/2019.

[3] Silhouette (clustering). https://en.wikipedia.org/wiki/Silhouette_(clustering). Último acceso: 28/12/2019.

[4] S Aranganayagi and K Thangavel. Clustering categorical data using silhouette coefficient as a relocating measure. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 2, pages 13–17. IEEE, 2007.

[5] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[6] N.H. Timm. *Applied Multivariate Analysis*. Springer Texts in Statistics. Springer New York, 2002.