Datos Masivos II

Descomposición CUR

Blanca Vázquez y Gibran Fuentes-Pineda

16 de agosto de 2022

Un poco de historia

- Surge en los años 1997-1998 por Stewart Stewart y Goreinov-Tyrtyshnikov-Zamarashkin.
- Este método se formaliza en el 2004 por Drineas-Kannan-Mahoney
- Surge como una alternativa a las descomposiciones deterministas para grandes volúmenes de datos.

¿Por qué surge la descomposición CUR?

Bolsa de palabras

	w_1	w_2	w_3	w_4	w_5	...	w_n
doc_1	0	0	1	0	0	0	1
doc_2	0	1	0	0	0	0	0
doc_3	0	0	0	1	0	0	1
doc_4	1	0	0	0	0	0	0
.....	...	...	...	...	...	...	...
doc_n	0	0	0	0	1	0	0

¡El reto de las matrices dispersas!

Descomposición de CUR

- Es un método de reducción, que se expresa en términos de un sub-conjunto de las variables originales.
- Es una alternativa aleatoria para SVD
- A diferencia de SVD que genera una descomposición exacta, CUR realiza una aproximación.

Descomposición de CUR

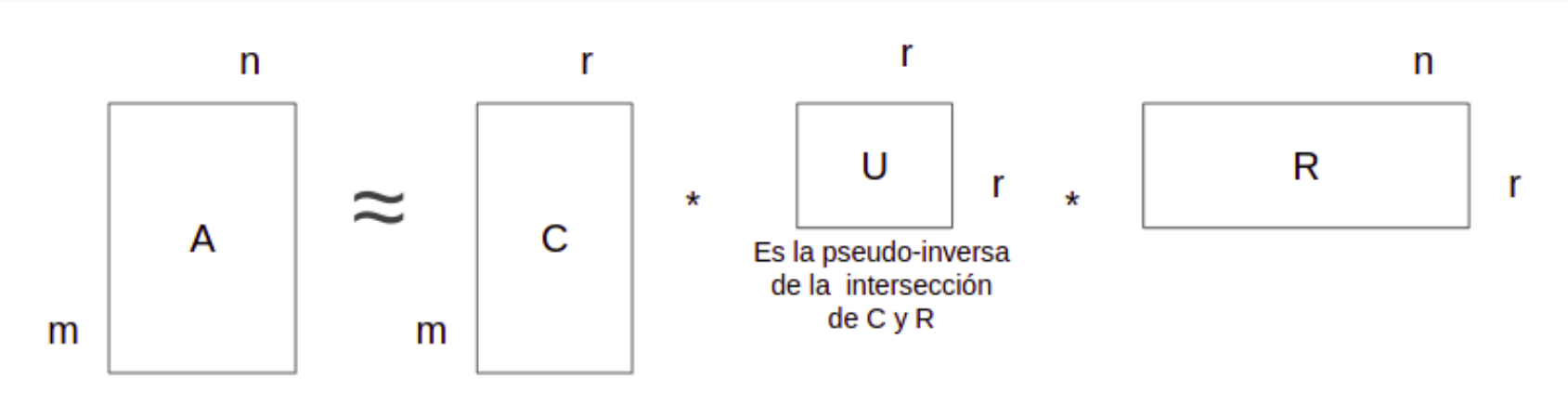
Dada una matriz A de *m* x *n*, se define la descomposición de CUR de A como:

$$M \approx CUR$$

Donde:

- *C* es una selección aleatoria de *r* columnas de A, y forma la matriz de *mxr*
- *R* es una selección aleatoria de *r* filas de A, y forma la matriz de *rxn*
- *U* es una matriz que se construye a partir de C y R
- r

Descomposición de CUR



Selección de filas y columnas

- La selección de filas y columnas se hace de manera aleatoria (se deben mantener las F y C más importantes).
- La medida de importancia es el cuadrado de la norma de Frobenius

$$f = \sum_{i,j} a_{ij}^2$$

- Se escala cada fila y columna seleccionada dividiendo sus elementos por la raíz cuadrada del número de veces esperado que esta fila y columna debería ser seleccionada, es decir $\sqrt{rp_i}$ o $\sqrt{rq_j}$.

Algoritmo para la selección de columnas

Muestreo de columnas (similar para filas)

Input: matrix A $\in \mathbb{R}^{mxn}$, sample size c

Output: $C_d \in \mathbb{R}^{mxc}$

1. for $x = 1 : n$ [column distribution]
    2. $P(x) = \Sigma_i A(i,x)^2 / \Sigma_{i,j} A(i,j)^2$
3. for $i = 1 : c$ [sample columns]
    4. Pick $j \in 1 : n$ based on distribution $P(x)$
    5. Compute $C_d(:,i) = A(:,j)/\sqrt{cP(j)}$

Cálculo de la matriz U

- Sea W la intersección de las columnas y filas muestreadas de C y R
- Calcular SVD para W, es decir: $W = X\Sigma Y^T$
- Entonces $U{=}W^+{=}Y\Sigma^+X^T$
  - Σ^+ son valores singulares distintos de cero: $\Sigma^+_{ii}{=}1/\Sigma_{ii}$
  - W^+ es la pseudo-inversa de Moore-Penrose

Ejercicio[1]

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

[1] Ejemplo tomado de Jure Leskovec, 2011.

Ejercicio

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

- La suma de los cuadrados de los elementos de $M = 243$

Ejercicio

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

- Cada una de las columnas de M, A y S tiene una norma cuadrada de Frobenius de $1^2 + 3^2 + 4^2 + 5^2 = 51$
- La probabilidad de cada una de las columnas para ser seleccionada es: $51/243 = 0.210$

Ejercicio

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

- Cada una de las columnas de C y T tiene una norma cuadrada de Frobenius de $4^2 + 5^2 + 2^2 = 45$
- La probabilidad de cada una de las columnas para ser seleccionada es: $45/243 = 0.185$

Ejercicio

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

- La probabilidad queda de la siguiente manera:
  P(M) = 0.210, P(A) = 0.210, P(S) = 0.210
  P(C) = 0.185, P(T) = 0.185

Ejercicio

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

- Para la fila 1 (Joe) $1^2 + 1^2 + 1^2 = 3$
- La probabilidad de que la fila sea seleccionada es $3/243 = 0.012$

Ejercicio

	Matrix	Alien	Star Wars	Casablanca	Titanic	Frobenius	Prob
Joe	1	1	1	0	0	3	0.12
Jim	3	3	3	0	0	27	0.111
John	4	4	4	0	0	48	0.198
Jack	5	5	5	0	0	75	0.309
Jill	0	0	0	4	4	32	0.132
Jenny	0	0	0	5	5	50	0.206
Jane	0	0	0	2	2	8	0.033
Frobenius	51	51	51	45	45		
Prob	0.210	0.210	0.210	0.185	0.185		

Selección de columnas

- La selección de columnas es aleatoria, sin embargo no es una probabilidad uniforme.
- Recordemos que la *jth* columna se selecciona con una probabilidad de q_j
- Cada columna de C se escoge independientemente de las columnas M, hay una probabilidad de escoger una columna más de una vez.
- Supongamos que r = 2 (no. columnas y filas a seleccionar)
- Cada columna seleccionada debe escalarse: dividiendo sus elementos entre $\sqrt{rq_j}$

Escalando columnas

$$Alien = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, Casablanca = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 5 \\ 2 \end{bmatrix}$$

Escalando columnas

$$Alien = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 5 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Calculamos $\sqrt{rq_2}$

Por lo tanto: $\sqrt{rq_2} = \sqrt{2 * 0.210} = 0.648$

Escalando columnas

$$Alien = \begin{bmatrix} 1/0.648 \\ 3/0.648 \\ 4/0.648 \\ 5/0.648 \\ 0/0.648 \\ 0/0.648 \\ 0/0.648 \end{bmatrix}, Alien_{esc} = \begin{bmatrix} 1.54 \\ 4.63 \\ 6.17 \\ 7.72 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$Alien_{esc}$ es la primera columna de la matriz C.

Escalando columnas

$$Casablanca = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 5 \\ 2 \end{bmatrix}$$

Calculamos $\sqrt{rq_4}$

Por lo tanto: $\sqrt{rq_4} = \sqrt{2 * 0.185} = 0.608$

Escalando columnas

$$Casablanca = \begin{bmatrix} 0/0.608 \\ 0/0.608 \\ 0/0.608 \\ 0/0.608 \\ 4/0.608 \\ 5/0.608 \\ 2/0.608 \end{bmatrix}, Casablanca_{esc} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 6.58 \\ 8.22 \\ 3.29 \end{bmatrix}$$

$Casablanca_{esc}$ es la segunda columna de la matriz C.

1er resultado: matriz C

$$C = \begin{bmatrix} 1.54 & 0 \\ 4.63 & 0 \\ 6.17 & 0 \\ 7.72 & 0 \\ 0 & 6.58 \\ 0 & 8.22 \\ 0 & 3.29 \end{bmatrix}$$

Selección de filas

- Supongamos que r = 2 (no. columnas y filas a seleccionar)
- Cada fila seleccionada debe escalarse: dividiendo sus elementos entre $\sqrt{rp_i}$

Selección de filas

$$Jenny = \begin{bmatrix} 0 & 0 & 0 & 5 & 5 \end{bmatrix}$$

$$Jack = \begin{bmatrix} 5 & 5 & 5 & 0 & 0 \end{bmatrix}$$

Escalando filas

$$Jenny = \begin{bmatrix} 0 & 0 & 0 & 5 & 5 \end{bmatrix}$$

Calculamos $\sqrt{rp_6}$

Por lo tanto: $\sqrt{rp_6} = \sqrt{2 * 0.206} = 0.642$

Escalando filas

$$Jenny = \begin{bmatrix} 0/0.642 & 0/0.642 & 0/0.642 & 5/0.642 & 5/0.642 \end{bmatrix}$$

$$Jenny_{esc} = \begin{bmatrix} 0 & 0 & 0 & 7.79 & 7.79 \end{bmatrix}$$

$Jenny_{esc}$ es la primera fila de la matriz R.

Escalando filas

$$Jack = \begin{bmatrix} 5 & 5 & 5 & 0 & 0 \end{bmatrix}$$

Calculamos $\sqrt{rp_4}$

Por lo tanto: $\sqrt{rp_4}$= $\sqrt{2 * 0.309}$ = 0.786

Escalando filas

$$Jack = \begin{bmatrix} 5/0.786 & 5/0.786 & 5/0.786 & 0/0.786 & 0/0.786 \end{bmatrix}$$

$$Jack_{esc} = \begin{bmatrix} 6.36 & 6.36 & 6.36 & 0 & 0 \end{bmatrix}$$

$Jack_{esc}$ es la segunda fila de la matriz R.

2do resultado: matriz R

$$R = \begin{bmatrix} 0 & 0 & 0 & 7.79 & 7.79 \\ 6.36 & 6.36 & 6.36 & 0 & 0 \end{bmatrix}$$

Construyendo la matriz U

- U es una matriz de r x r
- Para construir U, es necesario construir la matriz W.
- W es la intersección resultante entre filas y columnas que se usaron para construir C y R.
- W es una muestra de las filas y columnas de M con la mayor probabilidad
- Una vez construida W, se calcula su SVD, es decir: $W = X\Sigma Y^T$
- Cada elemento de la matriz Σ debe reemplazarse usando la pseudoinversa de Mooroe-Penrose
- Para obtener U, se calcula

$$U = Y(\Sigma^+)^2 X^T$$

Construyendo la matriz U

		Columnas para construir C				
		Matrix	Alien	Star Wars	Casa	Titanic
Filas para construir R	Jenny	0	0	0	5	5
	Jack	5	5	5	0	0

$$W = \begin{bmatrix} 0 & 5 \\ 5 & 0 \end{bmatrix}$$

Construyendo la matriz U

El siguiente paso es calcular la SVD de W

$$W = X\Sigma Y^T$$

Construyendo la matriz U

La SVD de $W = X\Sigma Y^T$:

$$W = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Construyendo la matriz U

Usando la matriz Σ, calculamos la pseudoinversa de Moore-Penrose[2] Σ^{+}

$$\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

- Cada elemento de la diagonal de Σ se reemplaza por $1/\sigma_i$
- Si el elemento de la diagonal es 0, se deja igual.

[2] Es una generalización de una matriz inversa, desarrollada en 1920

Construyendo la matriz U

Usando la matriz Σ, calculamos la pseudoinversa de Moore-Penrose Σ^+

$$\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}, \Sigma^+ = \begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix}$$

Construyendo la matriz U

Para calcular U, debemos calcular:

$$U = Y(\Sigma^{+})^{2}X^{T}$$

$$(X)\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, (\Sigma^{+})\begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix}, (Y^{T})\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Construyendo la matriz U

Para calcular U, debemos calcular:

$$U = Y(\Sigma^{+})^{2}X^{T}$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix}^{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Construyendo la matriz U

Para calcular U, debemos calcular:

$$U = Y(\Sigma^{+})^{2}X^{T}$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix}^{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/25 & 0 \\ 0 & 1/25 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} 1/25 & 0 \\ 0 & 1/25 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} 0 & 1/25 \\ 1/25 & 0 \end{bmatrix}$$

3er resultado: matriz U

$$U = \begin{bmatrix} 0 & 1/25 \\ 1/25 & 0 \end{bmatrix}$$

Descomposición CUR de M

Dada una matriz M de *m* x *n*, se define la descomposición de CUR de M como:

$$M \approx CUR$$

$$M \approx \begin{bmatrix} 1.54 & 0 \\ 4.63 & 0 \\ 6.17 & 0 \\ 7.72 & 0 \\ 0 & 6.58 \\ 0 & 8.22 \\ 0 & 3.29 \end{bmatrix} \begin{bmatrix} 0 & 1/25 \\ 1/25 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 7.79 & 7.79 \\ 6.36 & 6.36 & 6.36 & 0 & 0 \end{bmatrix}$$

Recordemos

- La descomposición CUR es un método de aproximación de la matriz original, construida a partir de un subconjunto de valores originales.
- Es un método de reducción de variables
- Es un método alternativo para SVD para matrices dispersas.

Aplicaciones

- Reconocimiento y análisis de imágenes
- Análisis término-documento
- Análisis de sistemas de recomendación
- Análisis de microarrays genómicos de ADN