Tema 2
Actividad 1

Autor
Juan José Méndez Torrero

Ejecución de FP-Growth

Introducción

Para la realización de esta actividad, se ha elegido el conjunto de datos Adult, el cual se puede encontrar dentro del repositorio UCI Machine Learning. Este conjunto de datos cuenta con 48842 instancias y con 14 atributos, tanto continuos como discretos. El objetivo de este conjunto de datos es determinar si una persona adulta sobrepasa el rango de ingresos de 50K dólares anuales.

Antes de la ejecución del algoritmo FP-Growth, se ha realizado un procesamiento del conjunto de datos, en el cual, se ha eliminado la columna "fnlwgt", la cual hacía referencia a un identificador.

Con respecto a los atributos continuos, se ha realizado un proceso de discretización para transformar esos atributos en discretos. Para ello, se ha realizado una discretización en tres rangos para los atributos "age", "education-num", "capital-gain", "capital-loss" y "hours-per-week".

Para la evaluación del algoritmo FP-Growth, se ha utilizado distintas configuraciones, las cuales resultarán en distintos patrones frecuentes para los distintos large-itemsets. Los valores de soporte mínimo que se han utilizado son: 0.1, 0.3, 0.5 y 0.7. Además, se hará un procesado de los resultados para obtener reglas que incluyan el atributo *class*, así, podríamos dar con alguna itemsets que nos pueda explicar la diferencia entre ingresos.

El código utilizado para sacar los resultados se puede encontrar aquí.

Soporte - 0.1

Tras la ejecución del algoritmo FP-Growth con un soporte mínimo de 0.1, se puede observar que el algoritmo encuentra 8216 patrones frecuentes, de los cuales, conseguimos patrones frecuentes con un soporte del 0.99, como se puede observar en la Figura 1. Además, con esta configuración hemos conseguido itemsets de longitud 10.

```
La máxima longitud del itemset encontrado es: 10
      support                                           itemsets
0    0.994902                    (capital-gain_[0.0, 33333.0))
1    0.955958                    (capital-loss_[0.0, 1452.0))
2    0.895854                 (native-country_ United-States)
3    0.854269                                  (race_ White)
4    0.759183                                 (class_ <=50K)
...       ...                                             ...
8211 0.104238  (capital-gain_[0.0, 33333.0), occupation_ Craf...
8212 0.103900  (capital-gain_[0.0, 33333.0), occupation_ Craf...
8213 0.105713  (capital-gain_[0.0, 33333.0), relationship_ Un...
8214 0.104238  (capital-loss_[0.0, 1452.0), relationship_ Unm...
8215 0.104115  (capital-gain_[0.0, 33333.0), capital-loss_[0....

[8216 rows x 2 columns]
```

Figura 1

Por otra parte, si sólo analizamos los itemsets que incluyan el atributo clase (<=50K, >50K), se puede observar que el número de patrones frecuentes encontrados es de 6158, como se puede observar en la Figura 2. De esta figura, se puede observar que el soporte máximo alcanzado es de 0.75 para la clase con ingresos menores de 50K, mientras que para la clase contraria, hemos obtenido un soporte del 0.24, con lo que se puede observar que dentro del conjunto de datos existe un número mayor de transacciones que cuentan con el ítem ***class <= 50K***.

	support	itemsets
0	0.759183	(class_ <=50K)
1	0.697052	(workclass_ Private)
2	0.240817	(class_ >50K)
3	0.758968	(capital-gain_[0.0, 33333.0), class_ <=50K)
4	0.738821	(class_ <=50K, capital-loss_[0.0, 1452.0))
...	...	...
6153	0.103501	(capital-gain_[0.0, 33333.0), race_ White, edu...
6154	0.101597	(race_ White, education-num_[6.0, 11.0), class...
6155	0.101505	(capital-gain_[0.0, 33333.0), education-num_[6...
...		
6156	0.100184	(class_ <=50K, occupation_ Adm-clerical)
6157	0.100184	(capital-gain_[0.0, 33333.0), occupation_ Adm-...

[6158 rows x 2 columns]

Figura 2

Soporte - 0.3

Tras la ejecución del algoritmo FP-Growth con un soporte mínimo de 0.3, se puede observar que el número de patrones frecuentes encontrados, disminuye drásticamente. Como se puede observar en la Figura 3, el número de patrones frecuentes encontrados es de 717, y con un total de 7 ítems en el conjunto de itemsets.

La máxima longitud del itemset encontrado es: 7

	support	itemsets
0	0.994902	(capital-gain_[0.0, 33333.0))
1	0.955958	(capital-loss_[0.0, 1452.0))
2	0.895854	(native-country_ United-States)
3	0.854269	(race_ White)
4	0.759183	(class_ <=50K)
..	...	...
712	0.312991	(class_ <=50K, marital-status_ Never-married)
713	0.318857	(capital-gain_[0.0, 33333.0), marital-status_ ...
714	0.312869	(capital-gain_[0.0, 33333.0), marital-status_ ...
715	0.305590	(class_ <=50K, marital-status_ Never-married, ...
716	0.305467	(capital-gain_[0.0, 33333.0), marital-status_ ...

[717 rows x 2 columns]

Figura 3

En cuanto a los patrones frecuentes sólo teniendo en cuenta que aparezca la clase, se puede observar que también disminuye drásticamente, de 6158 a 394, como se puede observar en la Figura 4.

	support	itemsets
0	0.759183	(class_ <=50K)
1	0.697052	(workclass_ Private)
2	0.758968	(capital-gain_[0.0, 33333.0), class_ <=50K)
3	0.738821	(class_ <=50K, capital-loss_[0.0, 1452.0))
4	0.675614	(class_ <=50K, native-country_ United-States)
..	...	...
389	0.301904	(capital-gain_[0.0, 33333.0), class_ <=50K, ca...
390	0.312991	(class_ <=50K, marital-status_ Never-married)
391	0.312869	(capital-gain_[0.0, 33333.0), marital-status_ ...
...		
392	0.305590	(class_ <=50K, marital-status_ Never-married, ...
393	0.305467	(capital-gain_[0.0, 33333.0), marital-status_ ...

[394 rows x 2 columns]

Figura 4

UCO ONLINE

Soporte - 0.5

Para la configuración de un soporte mínimo del 0.5, se puede observar en la Figura 5, que el número de patrones frecuentes encontrados es de 131. Y que el tamaño del mayor itemset encontrado es de 5.

```
La máxima longitud del itemset encontrado es: 5
      support                                           itemsets
0    0.994902                    (capital-gain_[0.0, 33333.0))
1    0.955958                    (capital-loss_[0.0, 1452.0))
2    0.895854                 (native-country_ United-States)
3    0.854269                                     (race_ White)
4    0.759183                                    (class_ <=50K)
..        ...                                               ...
126  0.543428  (capital-gain_[0.0, 33333.0), age_[17.0, 41.33...
127  0.524355  (age_[17.0, 41.333), capital-loss_[0.0, 1452.0...
128  0.522819  (capital-gain_[0.0, 33333.0), age_[17.0, 41.33...
129  0.516370  (capital-gain_[0.0, 33333.0), age_[17.0, 41.33...
130  0.508385  (capital-gain_[0.0, 33333.0), age_[17.0, 41.33...

[131 rows x 2 columns]
```

Figura 5

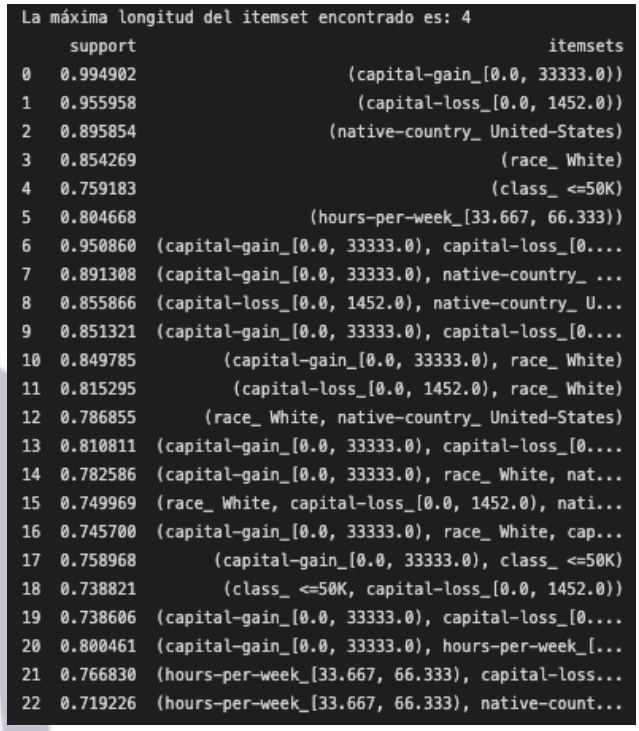
Por otro lado, en la Figura 6 se puede observar que el número de patrones encontrados que contentan el atributo clase, se ve disminuido hasta sólo llegar a 59 patrones.

```
    support                                           itemsets
0   0.759183                                    (class_ <=50K)
1   0.697052                                (workclass_ Private)
2   0.758968        (capital-gain_[0.0, 33333.0), class_ <=50K)
3   0.738821         (class_ <=50K, capital-loss_[0.0, 1452.0))
4   0.675614      (class_ <=50K, native-country_ United-States)
5   0.635688                          (class_ <=50K, race_ White)
6   0.584398    (class_ <=50K, hours-per-week_[33.667, 66.333))
7   0.738606  (capital-gain_[0.0, 33333.0), capital-loss_[0....
8   0.675430  (capital-gain_[0.0, 33333.0), native-country_ ...
...
56  0.512009  (class_ <=50K, capital-loss_[0.0, 1452.0), edu...
57  0.511855  (capital-gain_[0.0, 33333.0), class_ <=50K, ca...
58  0.508538                 (class_ <=50K, age_[17.0, 41.333))
59  0.508385  (capital-gain_[0.0, 33333.0), age_[17.0, 41.33...
```

Figura 6

Soporte - 0.7

Tras la ejecución del algoritmo FP-Growth con un soporte mínimo de 0.7, se puede observar en la Figura 7 que el número de patrones frecuentes encontrados sólo han sido 23, y que el itemsets más grande es de 4 ítems.

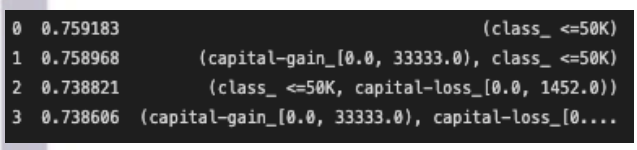


La máxima longitud del itemset encontrado es: 4

	support	itemsets
0	0.994902	(capital-gain_[0.0, 33333.0))
1	0.955958	(capital-loss_[0.0, 1452.0))
2	0.895854	(native-country_ United-States)
3	0.854269	(race_ White)
4	0.759183	(class_ <=50K)
5	0.804668	(hours-per-week_[33.667, 66.333))
6	0.950860	(capital-gain_[0.0, 33333.0), capital-loss_[0....
7	0.891308	(capital-gain_[0.0, 33333.0), native-country_ ...
8	0.855866	(capital-loss_[0.0, 1452.0), native-country_ U...
9	0.851321	(capital-gain_[0.0, 33333.0), capital-loss_[0....
10	0.849785	(capital-gain_[0.0, 33333.0), race_ White)
11	0.815295	(capital-loss_[0.0, 1452.0), race_ White)
12	0.786855	(race_ White, native-country_ United-States)
13	0.810811	(capital-gain_[0.0, 33333.0), capital-loss_[0....
14	0.782586	(capital-gain_[0.0, 33333.0), race_ White, nat...
15	0.749969	(race_ White, capital-loss_[0.0, 1452.0), nati...
16	0.745700	(capital-gain_[0.0, 33333.0), race_ White, cap...
17	0.758968	(capital-gain_[0.0, 33333.0), class_ <=50K)
18	0.738821	(class_ <=50K, capital-loss_[0.0, 1452.0))
19	0.738606	(capital-gain_[0.0, 33333.0), capital-loss_[0....
20	0.800461	(capital-gain_[0.0, 33333.0), hours-per-week_[...
21	0.766830	(hours-per-week_[33.667, 66.333), capital-loss...
22	0.719226	(hours-per-week_[33.667, 66.333), native-count...

Figura 7

Por otro lado, sólo ha encontrado 4 patrones frecuentes en los que se encuentra el atributo clase.



0	0.759183	(class_ <=50K)
1	0.758968	(capital-gain_[0.0, 33333.0), class_ <=50K)
2	0.738821	(class_ <=50K, capital-loss_[0.0, 1452.0))
3	0.738606	(capital-gain_[0.0, 33333.0), capital-loss_[0....

Figura 8

Conclusiones

Tras ejecutar el algoritmo FP-Growth con distintas configuraciones, se puede observar la importancia del soporte mínimo, ya que, como se ha visto, se pueden llegar a perder patrones a la hora de hacer una búsqueda de algún atributo en concreto. En este caso, el experto buscaría patrones que contengan el atributo *class*, ya que podría ser un patrón que pudiera explicar bastante bien la diferencia entre las clases.

Por ejemplo, de la última configuración utilizada (soporte mínimo de 0.7), se puede observar que se ha encontrado un patrón que es la siguiente:

{
'capital-gain_[0.0, 33333.0)',
'capital-loss_[0.0, 1452.0)',
'class_ <=50K'
}

De aquí, se podría decir que si el adulto está en el rango de ganancias [0.0, 33333.0) y tiene unas perdidas de [0.0, 1452.0), este adulto pertenecerá a la clase <=50K, con un soporte del 0.73.