Universidad de Córdoba
Escuela Politécnica Superior de Córdoba

Ingeniería Informática
Especialidad: Computación
Cuarto curso. Primer cuatrimestre

Introducción a la Minería de Datos.

Práctica 3: Reglas de asociación.

Victoria Pradas Laguna
github_victoriapl_MineriaDeDatos
i92prlav@uco.es

Curso académico 2022-2023
Córdoba, 4 de febrero de 2023

Índice

Índice de figuras

El objetivo de esta práctica es introducir los conceptos básicos de la obtención de reglas de asociación a partir de una lista de transacciones u otro tipo de ficheros de datos.

1. Utilizando el conjunto de datos *store_ data*

y ejecute el programa *assoc* para comprobar su funcionamiento. Intente interpretar las reglas obtenidas e indicar cuáles de ellas son importantes.

El programa *assoc* utiliza un algoritmo de reglas de asociación apriori para encontrar reglas de asociación en un conjunto de datos específico, *store_ data.csv*, que almacena las compras y transacciones de una tienda. Estas transacciones son obtenidas por el algoritmo como conclusión de que si un artículo es comprado con frecuencia en una transacción, también es probable que se compre a la vez otros artículos juntos.

Para que una regla sea considerada, se fija a partir del parámetro *min_ support*, siendo el valor mínimo de soporte. De igual forma, pero con el valor mínimo de confianza y lift sería *min_ confidence* y *min_ lift* correspondientemente. Este último parámetro, se refiere a la relación entre la probabilidad de comprar un conjunto de artículos juntos y la probabilidad de comprar cada artículo individualmente.

Tras ejecutar, se obtienen las reglas de asociación determinadas para ese conjunto de datos específico, *store_ data*y mostrando la relación entre diferentes items del dataset y mostrando las métricas de cada una de las reglas a partir de:

- Soporte: fracción de transiciones que contiene un conjunto de elementos en el conjunto de datos.
- Confianza: Medida de frecuencia con la que los artículos en Y aparecen en trasacciones que contienen X.
- Lift: Relación entre el soporte observado y el soporte esperado si los elementos uesen independientes.

La interpretación de cada una de las reglas son siempre iguales, por lo que se llevará a cabo la interpretación de la última: camarón entonces espaguetis.

```
================================
Rule: shrimp -> spaghetti
Support: 0.006
Confidence: 0.5232558139534884
Lift: 3.004914704939635
================================
```

La regla camarón entonces espagueti, tiene una métrica de **soporte** de 0.006, siendo un 0.6 % un valor menor que 1 %, indicando que de las transacciones existentes tan solo un 0.6 % que contienen ambos items o productos. De forma complementaria, la métrica de **confianza** es 0.5232, y por tanto existe un 52.32 % de que aquellas transacciones que contienen camarones también contengan espaguetis. Y por último, la métrica **lift** con un valor de 3 indicando que hay una fuerte asociación entre estos elementos, y por tanto cuanto mayor valor sea este, mayor asociación entre dichos elementos.

- La regla con soporte más alto es: ground beef → herb & pepper, indicando la regla más frecuente.
- La de confianza más alta: spagueti → ground beef.
- Lift más alto: pasta → escalope o nada, indicando la relación entre la frecuencia de aparición de una regla y la frecuencia de aparición de los items individuales.

Se debe tener en cuenta que por lo general con valores más altos tanto de soporte, confianza como lift se considerarán las reglas más importantes. Además, se debe tener en cuenta el ajuste de estos valores en función de los objetivos de cada análisis.

Por último se pueden observar los resultados del programa en modo de tabla en la Figura 1.

Como se ha dicho anteriormente, la regla será más importante cuanto mayor sea el valor de las métricas soporte, confianza y lift. En este caso las reglas más importantes podrían destacarse las siguientes:

```
ground_beef -> herb & pepper; support= 0.16
ground_beef -> frozen_vegetables & spaguetis; suport= 0.008
olive_oil -> whole wheat pasta: support=  0.008
```

index	items	support	ordered_statistics
0	frozenset({'mushroom cream sauce', 'escalope'})	0.005733333333333333	frozenset({'mushroom cream sauce'}),frozenset({'escalope'}),0.3006993006993073,3.790327319739845
1	frozenset({'pasta', 'escalope'})	0.005866666666666667	frozenset({'pasta'}),frozenset({'escalope'}),0.3728813559322034,4.700185158809287
2	frozenset({'ground beef', 'herb & pepper'})	0.016	frozenset({'herb & pepper'}),frozenset({'ground beef'}),0.3234501347708895,3.2915549671393096
3	frozenset({'tomato sauce', 'ground beef'})	0.005333333333333333	frozenset({'tomato sauce'}),frozenset({'ground beef'}),0.3773584905660376,3.8401474616626528
4	frozenset({'olive oil', 'whole wheat pasta'})	0.008	frozenset({'whole wheat pasta'}),frozenset({'olive oil'}),0.2714932126696833,4.130221288078346
5	frozenset({'pasta', 'shrimp'})	0.005066666666666666	frozenset({'pasta'}),frozenset({'shrimp'}),0.3220338983050848,4.514493901473151
6	frozenset({'frozen vegetables', 'shrimp', 'chocolate'})	0.005333333333333333	frozenset({'shrimp', 'chocolate'}),frozenset({'frozen vegetables'}),0.2962962962629634,3.1080031080831083
7	frozenset({'mushroom cream sauce', 'nan', 'escalope'})	0.005733333333333333	frozenset({'mushroom cream sauce'}),frozenset({'nan', 'escalope'}),0.3006993006993073,3.790327319739845,frozenset({'mushroom cream sauce', 'nan'}),frozenset({'escalope'}),0.3006993006993073,3.790327319739845
8	frozenset({'pasta', 'nan', 'escalope'})	0.005866666666666667	frozenset({'pasta'}),frozenset({'nan', 'escalope'}),0.3728813559322034,4.700185158809287,frozenset({'pasta', 'nan'}),frozenset({'escalope'}),0.3728813559322034,4.700185158809287
9	frozenset({'frozen vegetables', 'spaghetti', 'ground beef'})	0.005866666666666666	frozenset({'frozen vegetables', 'spaghetti'}),frozenset({'ground beef'}),0.3110047846889952,3.164906221394116
10	frozenset({'frozen vegetables', 'shrimp', 'mineral water'})	0.0072	frozenset({'mineral water', 'shrimp'}),frozenset({'frozen vegetables'}),0.3068181818181818,3.2163725365543547
11	frozenset({'frozen vegetables', 'spaghetti', 'tomatoes'})	0.006666666666666667	frozenset({'spaghetti', 'tomatoes'}),frozenset({'frozen vegetables'}),0.3184713375796178,3.3406084361449837
12	frozenset({'spaghetti', 'grated cheese', 'ground beef'})	0.005333333333333333	frozenset({'spaghetti', 'grated cheese'}),frozenset({'ground beef'}),0.3225806451612903,3.282706701096612
13	frozenset({'mineral water', 'ground beef', 'herb & pepper'})	0.006666666666666667	frozenset({'mineral water', 'herb & pepper'}),frozenset({'ground beef'}),0.399625,3.9751526458616161
14	frozenset({'ground beef', 'herb & pepper', 'nan'})	0.016	frozenset({'herb & pepper'}),frozenset({'ground beef', 'nan'}),0.3234501347708895,3.2915549671393096,frozenset({'nan', 'herb & pepper'}),frozenset({'ground beef'}),0.3234501347708895,3.2915549671393096
15	frozenset({'spaghetti', 'ground beef', 'herb & pepper'})	0.0064	frozenset({'spaghetti', 'herb & pepper'}),frozenset({'ground beef'}),0.3934426229508197,4.003825878061259
16	frozenset({'tomato sauce', 'ground beef', 'nan'})	0.005333333333333333	frozenset({'tomato sauce'}),frozenset({'ground beef', 'nan'}),0.3773584905660376,3.8401474616626528,frozenset({'tomato sauce', 'nan'}),frozenset({'ground beef'}),0.3773584905660376,3.8401474616626528
17	frozenset({'shrimp', 'spaghetti', 'ground beef'})	0.006	frozenset({'shrimp', 'ground beef'}),frozenset({'spaghetti'}),0.5232558139534884,3.954914470459635
18	frozenset({'nan', 'whole wheat pasta', 'olive oil'})	0.008	frozenset({'whole wheat pasta'}),frozenset({'nan', 'olive oil'}),0.2714932126696833,4.130221288078346,frozenset({'nan', 'whole wheat pasta'}),frozenset({'olive oil'}),0.2714932126696833,4.130221288078346
19	frozenset({'pasta', 'shrimp', 'nan'})	0.005066666666666666	frozenset({'pasta'}),frozenset({'shrimp', 'nan'}),0.3220338983050848,4.514493901473151,frozenset({'pasta', 'nan'}),frozenset({'shrimp'}),0.3220338983050848,4.514493901473151
20	frozenset({'frozen vegetables', 'shrimp', 'chocolate', 'nan'})	0.005333333333333333	frozenset({'shrimp', 'chocolate'}),frozenset({'frozen vegetables', 'nan'}),0.2962962962629634,3.1080031080831083,frozenset({'shrimp', 'chocolate', 'nan'}),frozenset({'frozen vegetables'}),0.2962962962629634,3.1080031080831083
21	frozenset({'frozen vegetables', 'spaghetti', 'ground beef', 'nan'})	0.008666666666666666	frozenset({'frozen vegetables', 'spaghetti'}),frozenset({'ground beef', 'nan'}),0.3110047846889952,3.164906221394116,frozenset({'frozen vegetables', 'spaghetti', 'nan'}),frozenset({'ground beef'}),0.3110047846889952,3.164906221394116
22	frozenset({'frozen vegetables', 'shrimp', 'nan', 'mineral water'})	0.0072	frozenset({'mineral water', 'shrimp'}),frozenset({'frozen vegetables', 'nan'}),0.3068181818181818,3.2163725365543547,frozenset({'mineral water', 'shrimp', 'nan'}),frozenset({'frozen vegetables'}),0.3068181818181818,3.2163725365543547
23	frozenset({'frozen vegetables', 'spaghetti', 'nan', 'tomatoes'})	0.006666666666666667	frozenset({'spaghetti', 'tomatoes'}),frozenset({'frozen vegetables', 'nan'}),0.3184713375796178,3.3406084361449837,frozenset({'spaghetti', 'nan', 'tomatoes'}),frozenset({'frozen vegetables'}),0.3184713375796178,3.3406084361449837
24	frozenset({'spaghetti', 'grated cheese', 'ground beef', 'nan'})	0.005333333333333333	frozenset({'spaghetti', 'grated cheese'}),frozenset({'ground beef', 'nan'}),0.3225806451612903,3.282706701096612,frozenset({'spaghetti', 'grated cheese', 'nan'}),frozenset({'ground beef'}),0.3225806451612903,3.282706701096612

Figura 1: Resultados reglas de asociación store_dataset

2. Usando los conjuntos *titanic.csv* y *bank-datafinal.arff* ejecute de nuevo el programa de generación de reglas, ordenando los resultados según su valor de lift. Interprete las reglas que se obtienen indicando su evaluación objetiva de interés.

2.1. Dataset *Titanic.csv*

Tras la ejecución del programa obtenemos de nuevo las métricas anteriores: soporte, confianza y lift. En este apartado, tendremos en cuenta la métrica lift y serán ordenados de mayor a menor; por lo que serán las primeras reglas las más importantes y más relevantes.

```
                      items   support  \
0          (29.125, 382652)  0.005612
1      (73.5, S.O.C. 14879)  0.005612
2       (0, 29.125, 382652)  0.005612
3  (0, 73.5, S.O.C. 14879)  0.005612
4  (S.O.C. 14879, 2, 73.5)  0.005612

                                  ordered_statistics
0  [((29.125), (382652), 1.0, 178.20000000000002)...
1  [((73.5), (S.O.C. 14879), 1.0, 178.20000000000...
2  [((29.125), (0, 382652), 1.0, 178.200000000000...
3  [((73.5), (0, S.O.C. 14879), 1.0, 178.20000000...
4  [((73.5), (2, S.O.C. 14879), 1.0, 178.20000000...
```

En este caso, como los resultados son números (que no indica que el valor de atributo sea numérico si no categórico) es más difícil de interpretar el conjunto de reglas. De todas formas, basándonos en las métricas podemos obtener las reglas más importantes cómo se muestra en el código anterior, esas son las reglas con mayor valor de lift y por tanto las que más destacan en este dataset según este apartado. Por otro lado, cabe destacar que a lo mejor en este dataset al no ser comercial como la del apartado uno la métrica más óptima podría ser soporte, habría que realizar un análisis más a fondo.

2.2. Dataset *Bankdatafinal.arff*

En el segundo caso, de igual forma se ordenarán de mayor valor de lift a menor, determinando la relación entre la frecuencia de aparición de una regla y la frecuencia de aparición de los items individuales.

```
                                          items   support  \
0           (FEMALE, 3, 52_max, 43759_max, RURAL)  0.008347
1       (FEMALE, 3, 52_max, 43759_max, NO, RURAL)  0.008347
2      (FEMALE, 3, YES, 52_max, 43759_max, RURAL)  0.008347
3  (FEMALE, 3, YES, 52_max, 43759_max, NO, RURAL)  0.008347
4               (3, 52_max, 43759_max, NO, RURAL)  0.011686

                                ordered_statistics
0  [((3, 43759_max), (RURAL, 52_max, FEMALE), 0.6...
1  [((3, 43759_max), (NO, RURAL, 52_max, FEMALE),...
2  [((3, 43759_max), (RURAL, 52_max, FEMALE, YES)...
3  [((3, 43759_max), (FEMALE, YES, 52_max, NO, RU...
4  [((3, 43759_max), (NO, RURAL, 52_max), 0.875, ...
```

Como se observa, ese sería el conjunto de reglas ordenados por lift y por tanto las reglas más importantes para este apartado. En diferencia al apartado anterior, al existir variables categóricas la interpretación es más intuitiva y podemos sacar información de las reglas más fácilmente.

Teniendo en cuenta la métrica lift como se ha hecho referencia en este apartado, las reglas más importantes son las obtenidas en el código anterior, pudiendo determinar información relevante como es el caso de que cuando una persona tiene tres hijos y sus ingresos máximos son de 43758, entonces generalmente es una mujer rural de edad no mayor a 52 años.

3. Seleccione al menos un nuevo conjunto de datos de los suministrados en los repositorios habituales. Para que se pueda usar el método *a priori* es necesario que los conjuntos no contentan variables numéricas

En este caso, se ha tenido en cuenta el conjunto de datos *Ballons.csv*, ya que sus variables son categóricas y será fácil de interpretar. Además tiene pocas instancias por lo que su tiempo de ejecución será menor que el de los anteriores.

El conjunto de datos tiene cinco atributos: el color del globo (amarillo o morado), el tamaño (pequeño o grande), proceso al que se le somete (estirar o sumergir), edad de la persona que le corresponde (niño o adulto) y finalmente si el globo está inflado o no con una variable de tipo booleana. En la Figura 2, se puede observar las cinco primeras instancias del conjunto de datos.

	Color	Size	Act	Age	Inflated
0	YELLOW	SMALL	STRETCH	CHILD	T
1	YELLOW	SMALL	DIP	ADULT	T
2	YELLOW	SMALL	DIP	CHILD	T
3	YELLOW	LARGE	STRETCH	ADULT	T
4	YELLOW	LARGE	STRETCH	CHILD	F

Figura 2: Globos Datasets

	ITEMS	SUPPORT	ORDERED_STATISTICS
0	(SMALL, CHILD, YELLOW, T)	[HTML]FFFFFF[HTML]212121 0.1333	[((CHILD, T), (SMALL, YELLOW), 1.0, 5.0), ((SMALL, YELLOW), (CHILD,T), 0.666, 5)]
1	(DIP, SMALL, YELLOW, T)	[HTML]FFFFFF[HTML]212121 0.1333	[((DIP, T), (SMALL, YELLOW), 1.0, 5.0), ((SMALL, YELLOW),(DIP,T), 0.666, 5)]
2	(LARGE, ADULT, STRETCH, T)	[HTML]FFFFFF[HTML]212121 0.1333	[((ADULT, STRETCH), (LARGE, T), 0.666, 5), ((LARGE,T), (ADULT,STRETCH), 1, 5)]
3	(T, ADULT, STRETCH, PURPLE)	[HTML]FFFFFF[HTML]212121 0.1333	[((ADULT, STRETCH),(LARGE,T), 0.666, 5), ((T, PURPLE),(ADULT,STRETCH), 1, 5)]
4	(SMALL, ADULT, YELLOW, DIP, T)	0.0667	[((DIP, T), (ADULT, SMALL, YELLOW), 0.5, 7.5), ((SMALL, YELLOW), (DIP,ADULT,T), 0.333, 5), ((DIP,ADULT,T),(SMALL, YELLOW), 1, 5, ((ADULT,SMALL, YELLOW)(DIP,T), 1, 7.5]

Tabla 1: Resultado reglas de asociación ordenada por la métrica confianza

	ITEMS	SUPPORT	ORDERED_STATISTICS
0	(SMALL, ADULT, YELLOW, DIP, T)	0.0667	[((DIP, T), (ADULT, SMALL, YELLOW), 0.5, 7.5), ((SMALL, YELLOW), (DIP,ADULT,T), 0.333, 5), ((DIP,ADULT,T),(SMALL, YELLOW), 1, 5, ((ADULT,SMALL, YELLOW)(DIP,T), 1, 7.5]
1	(SMALL, STRETCH, CHILD, YELLOW, T)	[HTML]FFFFFF[HTML]212121 0.0667	[((CHILD, T),(SMALL,STRETCH, YELLOW), 0.5, 7.5), ((SMALL, YELLOW),(CHILD, STRETCH, T), 0.333, 5), ((CHILD, STRETCH, T), (SMALL, YELLOW), 1, 5), ((SMALL, STRETCH, YELLOW), (CHILD,T), 1, 7.5)]
2	(LARGE, ADULT, STRETCH, T)	[HTML]FFFFFF[HTML]212121 0.1333	[((ADULT, STRETCH), (LARGE, T), 0.666, 5), ((LARGE,T), (ADULT,STRETCH), 1, 5)]
3	(T, ADULT, STRETCH, PURPLE)	[HTML]FFFFFF[HTML]212121 0.1333	[((ADULT, STRETCH),(LARGE,T), 0.666, 5), ((T, PURPLE),(ADULT,STRETCH), 1, 5)]
4	(SMALL, CHILD, YELLOW, T)	[HTML]FFFFFF[HTML]212121 0.1333	[((CHILD, T), (SMALL, YELLOW), 1.0, 5.0), ((SMALL, YELLOW), (CHILD,T), 0.666, 5)]

Tabla 2: Resultado reglas de asociación ordenada por la métrica lift.

4. Seleccione las cinco mejores reglas usando las medidas de confianza y lift. Compare las reglas obtenidas y comente qué información puede obtener de algunas de las reglas.

Aplicando el método a priori haciendo uso del método *sorted* para realizar la ordenación de mayor a menor valor de la métrica **confianza**, obtenemos los siguientes resultados:

Tras la obtención de la Tabla1 que determina las cinco mejores reglas ordenadas por la métrica confianza y 2 las cinco mejores ordenadas por la métrica lift. Podemos observar que los valores y el orden es parecido.

El resultado total obtiene una tabla con 14 filas, por lo que obteniendo las 5 mejores obtenemos las nombradas anteriormente. El orden, como se observa en las Tablas 1 2 es similar, y destacan las mismas reglas, por lo que esto significa que son las más importantes y las que más información aportan para posteriormente poder predecir más información.

Entre las reglas que se han obtenido, podemos sacar la siguiente informa-

ción:

- (SMALL, YELLOW) → (CHILD,T). Esta regla indica que si existe globo pequeño y amarillo entonces es de un niño y está inflado.
- (SMALL, YELLOW) → (DIP,T). Esta regla de forma similar a la anterior, indica que si existe un globo pequeño y amarillo, entonces ha sido sumergido en agua y está inflado.
- (ADULT, STRETCH) → (LARGE,T). Si la persona que tiene el globo es un adulto y estira de éste, entonces el globo suele ser largo y estar inflado o viceversa.

Estos items, son ejemplos de la información obtenida del resultado de las reglas. Además hay que destacar que las reglas no se cumplen siempre si no que hay que atenerse a las métricas estudiadas y que son *generalizaciones* obtenidas por el estudio de la ocurrencia de los items.