Practica 5: Reglas de asociación

Profesor: Lauro Reyes Cocoletzi

Nancy Galicia y Daphne González
{ngaliciac2100@alumno.ipnx.mx y dgonzalezc2104@alumno.ipnx.mx

UPIIT: Unidad Profesional Interdisciplinaria en Ingeniería Campus Tlaxcala Instituto Politécnico Nacional, Tlaxcala, Tlaxcala, México 9000

Ingeniera en Inteligencia Artificial

06 de noviembre del 2023

***Resumen*— La minería de datos es considerada una herramienta para extraer conocimiento en grandes volúmenes de información. Uno de los análisis realizados en minería de datos son las reglas de asociación, cuyo propósito es buscar coocurrencias entre los registros de un conjunto de datos. Su principal aplicación se encuentra en el análisis de canasta de mercado, donde se establecen criterios para la toma de decisiones a partir del comportamiento de compra de los clientes.**

Palabras clave* — *Reglas de asociación, antecedente, consecuente

I. Marco Teorico

A. Reglas de asociación

Las reglas de asociación son una técnica de minería de datos que permite descubrir patrones y relaciones ocultas, siendo este su principal objetivo.

1. Antecedente (Antecedent): Es un conjunto de elementos o ítems que se consideran como el punto de partida o la condición en una regla.
2. Consecuente (Consequent): Es un conjunto de elementos que se consideran como el resultado o la conclusión en una regla.

Cada regla de asociación se expresa en la forma "Si {antecedente} entonces {consecuente}", y se asocia con un nivel de soporte y confianza. El soporte indica cuántas veces aparece la regla en el conjunto de datos, y la confianza mide la probabilidad de que la regla sea válida.

B. Algoritmo Apiori

Se trata de uno de los algoritmos más populares para el descubrimiento de reglas de asociación, se basa en el principio del "apoyo descendente", que establece que si un conjunto de elementos (ítems) es frecuente, entonces todos sus subconjuntos también lo son. El proceso del algoritmo Apriori se divide en varios pasos:

1. Generación de conjuntos de elementos frecuentes: El algoritmo encuentra todos los elementos individuales (ítems) que tienen un soporte mayor o igual a un umbral predefinido. Genera conjuntos más grandes de elementos candidatos (combinando elementos individuales) y verifica su soporte en el conjunto de datos. Esto continua hasta que ya no se pueden generar conjuntos de elementos frecuentes.
2. Generación de reglas de asociación: Con los conjuntos de elementos frecuentes, el algoritmo Apriori genera reglas de asociación a partir de estos. Las reglas se crean combinando los elementos del conjunto frecuente en antecedentes y consecuentes posibles y calculando su soporte y confianza.
3. Selección de reglas significativas: Se aplican umbrales de soporte y confianza para seleccionar las reglas más interesantes o significativas.

Ventajas del Algoritmo Apriori

- Es un algoritmo fácil de implementar y de entender.
- Se puede utilizar en conjuntos de elementos grandes.

Desventajas del Algoritmo Apriori

- A veces, es necesario encontrar un gran número de reglas candidatas, lo que puede resultar caro desde el punto de vista informático.
- El cálculo del soporte también es caro porque tiene que recorrer toda la base de datos.

C. Fp growth

El FP-Growth (Frequent Pattern growth) deriva del "a-priori" y es un algoritmo que se caracteriza por ser muy eficiente y además escalable, es decir, se puede usar para grandes volúmenes de datos con un orden razonablemente bajo.

Se usa para encontrar conjuntos de patrones frecuentes en bases de datos sin generar candidatos. El algoritmo crea un árbol (FP-tree) de patrones frecuentes en el conjunto de datos.

Opera contando la frecuencia de cada item, cada par item, frecuencia es un candidato. Luego compara esto con el mínimo y lo ordena descendente para poder tratarlo.

Finalmente compara esta lista ordenada con la lista inicial, si el elemento está en la tabla inicial lo pasa a una tercera tabla de resultados que contiene los elementos más frecuentes.

D. ONE HOT ENCODING

Se trata de una forma sencilla de asignar valores a las diferentes categorías de una variable categorica. en esta estrategia se crea una columna binaria ara cada valor único que exista en la variable categórica que esta codificando y marcar con un 1 la columna correspondiente al valor presente en cada registro.

E. DATASET

CONJUNTO DE DATOS 1

Transacción				
1	A	B	C	D
2	A	B	D	
3	A	B		
4	B	C	D	
5	B	C		
6	C	D		
7	B	D		

CONJUNTO DE DATOS 2

Transacción				
1	I1	I2	I3	
2	I2	I3	I4	
3	I4	I5		
4	I1	I2	I4	
5	I1	I2	I3	I5
6	I1	I2	I3	I4

CONJUNTO DE DATOS PROPIO

Closet DataSet

1	Camiseta	Jeans	Vestido	Sudadera	
2	Chaqueta	Gorro	Vestido	Botas	Falda
3	Gabardina	Pantalón	Camiseta		
4	Falda	Jeans	Chaqueta	Gorro	
5	Vestido	Botas	Sudadera		
6	Pantalón	Camiseta	Chaqueta	Gorro	
7	Falda	Sudadera	Jeans	Chaqueta	
8	Gabardina	Botas	Vestido	Chaqueta	Camiseta
9	Sudadera	Chaqueta	Pantalón	Falda	Gorro
10	Jeans	Vestido	Botas	Chaqueta	Camiseta
11	Vestido	Camiseta	Chaqueta		
12	Pantalón	Sudadera	Gorro		
13	Chaqueta	Falda	Jeans	Botas	
14	Camiseta	Vestido	Gorro	Sudadera	Pantalón
15	Botas	Falda	Chaqueta	Jeans	

II. DESARROLLO

Item-Set con soporte mínimo en 2

CONJUNTO DE DATOS 1

{ 'A' : 3, 'B': 6, 'C' = 4, 'D': 5]

Frecuencia 1-itemset	['A', 'B', 'C', 'D']
Frecuencia 2-itemset	['AB', 'AD', 'BC']
Frecuencia 3-itemset	['ABD', 'BCD']

Mínimo conteo de soporte: 3

Todas las frecuencias de los itemsets:
[{'A'}, {'B', 'A'}, {'C'}, {'D', 'C'}, {'B', 'C'}, {'D'}, {'D', 'B'}, {'B'}]

CONJUNTO DE DATOS 2

{'I1': 4, 'I2': 5, 'I3': 4, 'I4': 4, 'I5': 2}

Frecuencia 1-itemset	['I1', 'I2', 'I3', 'I4', 'I5']
Frecuencia 2-itemset	['I1I2', 'I1I3', 'I1I4', 'I2I3', 'I2I4', 'I3I4']
Frecuencia 3-itemset	['I1I2I3', 'I1I2I4', 'I2I3I4']

Mínimo conteo de soporte: 3

Todas las frecuencias de los itemsets:
[{'I1'}, {'I2', 'I1'}, {'I3'}, {'I1', 'I3'}, {'I2', 'I1', 'I3'}, {'I2', 'I3'}, {'I4'}, {'I4', 'I2'}, {'I2'}]

CONJUNTO DE DATOS PROPIO

A. CODIFICACIÓN DE DATOS

Para la aplicación del algoritmo de a priori, se convirtió el dataset a una lista booleana one hot encoded y de esta manera poder identificar con mayor facilidad los patrones que existen en la posesión de prendas de cada armario representado dentro del dataset.

Botas	**Camiseta**	**Chaqueta**	**Falda**	**Gabardina**
0	True	0	0	0
True	0	True	True	0
0	True	0	0	True
0	0	True	True	0

True	0	0	0	0
0	True	True	0	0
0	0	True	True	0
True	True	True	0	True
0	0	True	True	0
True	True	True	0	0
0	True	True	0	0
0	0	0	0	0
True	0	True	True	0
0	True	0	0	0
True	0	True	True	0

Gorro	Jeans	Pantalón	Sudadera	Vestido
0	True	0	True	True
True	0	0	0	True
0	0	True	0	0
True	True	0	0	0
0	0	0	True	True
True	0	True	0	0
0	True	0	True	0
0	0	0	0	True
True	0	True	True	0
0	True	0	0	True
0	0	0	0	True
True	0	True	True	0
0	True	0	0	0
True	0	True	True	True
0	True	0	0	0

B. APLICACIÓN DEL ALGORITMO A PRIORI

Comenzamos por determinar el soporte mínimo para las transacciones como un 20% garantizando que los conjuntos de elementos considerados sean lo suficientemente frecuentes como para ser significativos y útiles en la identificación de patrones y reglas de asociación relevantes en el conjunto de datos.

	SUPPORT	ITEMSETS
0	0.4	(Botas)
1	0.466667	(Camiseta)
2	0.666667	(Chaqueta)
3	0.4	(Falda)
4	0.4	(Gorro)
5	0.4	(Jeans)
6	0.333333	(Pantalón)
7	0.4	(Sudadera)
8	0.466667	(Vestido)
9	0.333333	(Botas, Chaqueta)
10	0.2	(Botas, Falda)
11	0.2	(Botas, Jeans)
12	0.266667	(Botas, Vestido)
13	0.266667	(Camiseta, Chaqueta)
14	0.2	(Camiseta, Pantalón)
15	0.333333	(Camiseta, Vestido)
16	0.4	(Chaqueta, Falda)
17	0.266667	(Chaqueta, Gorro)
18	0.333333	(Chaqueta, Jeans)
19	0.266667	(Chaqueta, Vestido)
20	0.2	(Gorro, Falda)
21	0.266667	(Falda, Jeans)
22	0.266667	(Gorro, Pantalón)
23	0.2	(Sudadera, Gorro)
24	0.2	(Sudadera, Pantalón)
25	0.2	(Sudadera, Vestido)
26	0.2	(Botas, Chaqueta, Falda)
27	0.2	(Botas, Chaqueta, Jeans)
28	0.2	(Botas, Chaqueta, Vestido)
29	0.2	(Camiseta, Chaqueta, Vestido)
30	0.2	(Gorro, Chaqueta, Falda)
31	0.266667	(Chaqueta, Falda, Jeans)
32	0.2	(Sudadera, Gorro, Pantalón)

Para la segunda parte se establece que la confianza de las reglas de asociación resultantes fuera mayor a 0.7 para que las reglas de asociación que se identifiquen sean altamente confiables y a descartar aquellas reglas que podrían surgir debido a coincidencias aleatorias

ANTECEDENTS	CONSEQUENTS	SUPPORT	CONFIDENCE	LIFT
(Botas)	(Chaqueta)	0.333333	0.833333	1.25
(Falda)	(Chaqueta)	0.4	1	1.5
(Jeans)	(Chaqueta)	0.333333	0.833333	1.25
(Pantalón)	(Gorro)	0.266667	0.8	2
(Botas, Falda)	(Chaqueta)	0.2	1	1.5
(Botas, Jeans)	(Chaqueta)	0.2	1	1.5
(Gorro, Falda)	(Chaqueta)	0.2	1	1.5
(Chaqueta, Jeans)	(Falda)	0.266667	0.8	2
(Falda, Jeans)	(Chaqueta)	0.266667	1	1.5
(Sudadera, Gorro)	(Pantalón)	0.2	1	3
(Sudadera, Pantalón)	(Gorro)	0.2	1	2.5

RESULTADOS

RESULTADOS DE LAS REGLS DE ASOCIACIÓN DEL CONJUNTO 1

Si {A} entonces {B} (Confianza = 100%)
Si {C} entonces {B} (Confianza = 100%)
Si {D} entonces {B} (Confianza = 100%)
Si {C} entonces {D} (Confianza = 100%)
Si {D} entonces {C} (Confianza ≈ 75%)

Resultados de las reglas de asociación conjunto 2

Si {I1} entonces {I2} (Confianza = 100%)
Si {I3} entonces {I2} (Confianza = 100%)
Si {I1, I3} entonces {I2} (Confianza = 100%)
Si {I2} entonces {13} (Confianza ≈ 80%)

Resultados de las reglas de asociación del dataset selecionado

falda → chaqueta
Suporte: 0.4
Lift: 1.5
Confianza: 1

Debido a que teneos una confianza de 1, quiere decir que en donde exista una falda se encuentra una chaqueta.

Botas, falda → Chaqueta
Suporte: 0.2
Lift: 1.5
Confianza: 1

Debido a que teneos una confianza de 1, quiere decir que en donde exista una falda se encuentra una chaqueta.

Botas, Jeans → Chaqueta
Suporte: 0.2
Lift: 1.5
Confianza: 1

Debido a que teneos una confianza de 1, quiere decir que en donde exista botas y jeans se encuentra una chaqueta.

Gorro, Falda → Chaqueta
Suporte: 0.2
Lift: 1.5
Confianza: 1

Debido a que teneos una confianza de 1, quere decir que en donde exista gorro y falda se encuentra una chaqueta.

Falda, Jeans → Chaqueta
Suporte: 0.26666
Lift: 1.5
Confianza: 1

Debido a que teneos una confianza de 1, quiere decir que en donde exista una falda y jeans se encuentra una chaqueta.

Sudadera, Gorro → Pantalón
Suporte: 0.2
Lift: 3
Confianza: 1

Debido a que teneos una confianza de 1, quiere decir que en donde exista una sudadera y gorro se encuentra un pantalón.

Sudadera, Pantalón → Gorro
Suporte: 0.2
Lift: 3
Confianza: 1

Debido a que teneos una confianza de 1, quiere decir que en donde exista una sudadera y pantalón se encuentra un gorro.

Al observar las reglas de asociación con una confianza del 100% en el conjunto de datos, se destaca que la presencia de una chaqueta es la consecuencia más común en los armarios designados. Este hallazgo cobra sentido si consideramos el filtro inicial, que revela que las chaquetas tienen el índice de soporte más alto, alcanzando un 66%. Esto indica que más de la mitad de los armarios incluyen una chaqueta.

Además, es crucial tener en cuenta las otras tres reglas de asociación reveladas en los resultados. Aunque no alcanzan el umbral del 100%, no deberíamos descartarlas por completo. En conjuntos de datos más extensos, es posible que no se encuentren reglas de asociación con una certeza absoluta, y estas reglas podrían proporcionar información valiosa sobre patrones de asociación relevantes.

I. Conclusiones

Las reglas de asociación y los algoritmos Apriori y FP-Growth son herramientas fundamentales en el campo de la minería de datos, ya que permiten descubrir relaciones y patrones ocultos dentro de un conjunto de datos.

Como observamos, el algoritmo Apriori se basa en la generación de candidatos y posteriormente descarta aquellos que no cumplen con un umbral de soporte mínimo. Esto lo hace más eficiente en conjuntos de datos pequeños, ya que el proceso de generación de candidatos puede volverse costoso en términos de tiempo y recursos computacionales. Sin embargo, puede no ser la mejor opción en conjuntos de datos más grandes debido a su complejidad computacional.

Por otro lado, el algoritmo FP-Growth utiliza una estructura de árbol condicional para reducir la exploración de candidatos, lo que lo hace más eficiente, especialmente en conjuntos de datos con elementos dispersos o con mayor cantidad de datos. En este caso, FP-Growth puede no ser la mejor opción debido al tamaño

de nuestras muestras ya que este algoritmo está diseñado especialmente para reducir significativamente el tiempo de procesamiento en comparación con Apriori.

Sin embargo, de cada algoritmo debemos comprender la función y descripción de cada una de las características para de esta manera seleccionar las reglas de asociación que más nos ayuden a lograr nuestros objetivos. Un ejemplo de esto es que en el algoritmo a priori no solo es importante considerar el valor en solitario de la confianza de la regla ya que muchas veces el Lift score nos puede dar una mayor noción de que tan útil será dicha regla para describir al conjunto de los datos.

Así pues, comprender el funcionamiento de cada algoritmo, así como analizar sus ventajas y desventajas en función de las características de un conjunto de datos específico, permite seleccionar el algoritmo que mejor se adapte a las necesidades y los resultados deseados. Por ejemplo, si se trabaja con un conjunto de datos pequeño, Apriori puede ser una elección adecuada, mientras que, en conjuntos de datos más grandes y dispersos, FP-Growth puede ser la opción preferida debido a su eficiencia en el procesamiento de datos. La elección del algoritmo dependerá de las características y objetivos de cada tarea de minería de datos.

II. BIBLIOGRAFIA

[1] Colaboradores de los proyectos Wikimedia. “Reglas de asociación - Wikipedia, la enciclopedia libre”. Wikipedia, la enciclopedia libre. Accedido el 3 de noviembre de 2023. [En línea]. Disponible: https://es.wikipedia.org/wiki/Reglas_de_asociación

[2] “Algoritmo Apriori - Teoría - 🤖 Aprende IA”. 🤖 Aprende IA. Accedido el 3 de noviembre de 2023. [En línea]. Disponible: https://aprendeia.com/algoritmo-apriori/

[3] “IBM Documentation”. IBM in Deutschland, Österreich und der Schweiz | IBM. Accedido el 3 de noviembre de 2023. [En línea]. Disponible: https://www.ibm.com/docs/es/spss-modeler/saas?topic=nodes-association-rules

[4] “Regresión Lineal en español con Python”. Aprende Machine Learning. Accedido el 3 de noviembre de 2023. [En línea]. Disponible: https://www.aprendemachinelearning.com/regresion-lineal-en-espanol-con-python/