

DATA MINING: REGLAS DE ASOCIACIÓN ASIMÉTRICAS

LABORATORIO DE MODELACIÓN II

GUSTAVO ARCAYA

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARIA
DEPARTAMENTO DE MATEMÁTICAS

26-07-2021

Trabajo propuesto por:

Sebastian Torres, Head of Merchandising Analytics, Walmart Chile

- 1 Definiciones Iniciales
- 2 Reglas de Asociación
 - Reglas Asimétricas
- 3 Implementación
- 4 Resultados
- 5 Otros enfoques
 - Propuestas finales
 - Trabajos futuros
 - Trabajos futuros

DEFINICIONES INICIALES

Consideraremos $I = \{i_1, \dots, i_n\}$ un conjunto de m variable binarias, que denominamos **items**

Consideraremos $I = \{i_1, \dots, i_n\}$ un conjunto de m variable binarias, que denominamos **items** y sea $T = \{t_1, \dots, t_n\}$ el conjunto de **transacciones** que obtenemos desde nuestra base de datos

Consideraremos $I = \{i_1, \dots, i_n\}$ un conjunto de m variable binarias, que denominamos **items** y sea $T = \{t_1, \dots, t_n\}$ el conjunto de **transacciones** que obtenemos desde nuestra base de datos, donde cada transacción $t_i \in T$ se identifica de forma única y corresponde un subconjunto de I , y se definen mediante un vector binario de la forma

$$\begin{cases} t(k) = 1, \text{ si } i_k \in t \\ t(k) = 0, \text{ en otro caso.} \end{cases}$$

REGLAS DE ASOCIACIÓN

Una regla de asociación se define como una “implicancia” de la forma $X \Rightarrow Y$, para $X, Y \subset I$, tal que los conjuntos X, Y no son disjuntos.

Podemos definir el **soporte** de un conjunto $X \subset I$ como $[4]^1$

Soporte

Para $X \subset I$, definimos el soporte de X como su frecuencia relativa, es decir

$$\text{supp}(X) = \frac{|\{t \in T : X \subset t\}|}{|T|} = \mathbb{P}(x) \in [0, 1]$$

¹P. Tan, “Selecting the right objective measure for association analysis.”

A partir del soporte $[3]^2$ podemos definir otras reglas entre items que resultan de interés para la toma de decisiones comerciales: promociones, localización en tienda y precios de los productos

²M. Hahsler, "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules."; GitHub

A partir del soporte [3]² podemos definir otras reglas entre ítems que resultan de interés para la toma de decisiones comerciales: promociones, localización en tienda y precios de los productos, entre ellas, definiremos

Confidence

Para X, Y itemsets dados, definimos

$$\text{conf}(X \Rightarrow Y) := \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \in [0, 1]$$

²M. Hahsler, "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules."; GitHub

A partir del soporte [3]² podemos definir otras reglas entre ítems que resultan de interés para la toma de decisiones comerciales: promociones, localización en tienda y precios de los productos, entre ellas, definiremos

Confidence

Para X, Y itemsets dados, definimos

$$\text{conf}(X \Rightarrow Y) := \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \in [0, 1]$$

Confidence corresponde a la proporción de transacciones que contienen Y dentro del conjunto de transacciones conteniendo X .

²M. Hahsler, "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules."; GitHub

Como antes, utilizaremos el soporte para la siguiente definición

Lift

Para X, Y itemsets dados, definimos el Lift como

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} = \frac{\mathbb{P}(Y|X)}{\mathbb{P}(Y)} = \frac{\mathbb{P}(X \cap Y)}{\mathbb{P}(X)\mathbb{P}(Y)} \in [0, \infty)$$

Como antes, utilizaremos el soporte para la siguiente definición

Lift

Para X, Y itemsets dados, definimos el Lift como

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} = \frac{\mathbb{P}(Y|X)}{\mathbb{P}(Y)} = \frac{\mathbb{P}(X \cap Y)}{\mathbb{P}(X)\mathbb{P}(Y)} \in [0, \infty)$$

Lift mide el ratio de aparición real del conjunto $X \cup Y$ respecto a la aparición teórica en caso de ser independientes.

¿POR QUÉ BUSCAR REGLAS ASIMÉTRICAS?

a) Establecer **causalidad** entre itemsets, [1]³.

³R. Agrawal, "Mining Association Rules between sets of item in Large Databases"

¿POR QUÉ BUSCAR REGLAS ASIMÉTRICAS?

- a) Establecer **causalidad** entre itemsets, [1]³.
- b) Mejora la toma de decisiones, ahorrando tiempo y potencialmente dinero a compradores.

³R. Agrawal, "Mining Association Rules between sets of item in Large Databases"

¿POR QUÉ BUSCAR REGLAS ASIMÉTRICAS?

- a) Establecer **causalidad** entre itemsets, [1]³.
- b) Mejora la toma de decisiones, ahorrando tiempo y potencialmente dinero a compradores.
- c) Obtención de una “jerarquía” u “ordenamiento” de productos según su impacto.

³R. Agrawal, “Mining Association Rules between sets of item in Large Databases”

Collective Strength

Para X, Y dos itemsets dados,

$$S(X) = \frac{\mathbb{P}(X \cap Y) + \mathbb{P}(\bar{Y}|\bar{X})}{\mathbb{P}(X)\mathbb{P}(Y) + \mathbb{P}(\bar{X})\mathbb{P}(\bar{Y})} \in [0, \infty]$$

Collective Strength

Para X, Y dos itemsets dados,

$$S(X) = \frac{\mathbb{P}(X \cap Y) + \mathbb{P}(\bar{Y}|\bar{X})}{\mathbb{P}(X)\mathbb{P}(Y) + \mathbb{P}(\bar{X})\mathbb{P}(\bar{Y})} \in [0, \infty]$$

Interpretamos este indicador como,

$$\begin{cases} S(X) = 0, X \text{ e } Y \text{ negativamente correlacionados,} \\ S(X) \approx 1, X \text{ e } Y \text{ se portan como estadísticamente independiente,} \\ S(X) \approx \infty, X \text{ e } Y \text{ están perfectamente correlacionados.} \end{cases}$$

Desde [4] y [3] se propone el siguiente indicador, verificaremos la direccionalidad de este:

Conviction

Sean X, Y datasets dados, definimos

$$\text{conviction}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)} = \frac{\mathbb{P}(X)\mathbb{P}(\bar{Y})}{\mathbb{P}(X \cap \bar{Y})} \in [0, \infty)$$

Desde [4] y [3] se propone el siguiente indicador, verificaremos la direccionalidad de este:

Conviction

Sean X, Y datasets dados, definimos

$$\text{conviction}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)} = \frac{\mathbb{P}(X)\mathbb{P}(\bar{Y})}{\mathbb{P}(X \cap \bar{Y})} \in [0, \infty)$$

Conviction compara la probabilidad de aparición de X sin Y respecto a si estos sucesos fueran dependientes.

IMPLEMENTACIÓN

Comenzamos con la forma en que se almacena la información de las transacciones

Comenzamos con la forma en que se almacena la información de las transacciones

```
transactions.head(15)
```

	Visit_date	Visit_nbr	Category
0	2017-01-02	700200000	3
1	2017-01-02	700200001	126
2	2017-01-02	700200001	130
3	2017-01-02	700200002	127
4	2017-01-02	700200002	38
5	2017-01-02	700200002	130
6	2017-01-02	700200002	15
7	2017-01-02	700200002	125
8	2017-01-02	700200003	37
9	2017-01-02	700200004	119
10	2017-01-02	700200004	165
11	2017-01-02	700200004	117
12	2017-01-02	700200005	101
13	2017-01-02	700200006	353
14	2017-01-02	700200007	74

Comenzamos con la forma en que se almacena la información de las transacciones

```
transactions.head(15)
```

	Visit_date	Visit_nbr	Category
0	2017-01-02	700200000	3
1	2017-01-02	700200001	126
2	2017-01-02	700200001	130
3	2017-01-02	700200002	127
4	2017-01-02	700200002	38
5	2017-01-02	700200002	130
6	2017-01-02	700200002	15
7	2017-01-02	700200002	125
8	2017-01-02	700200003	37
9	2017-01-02	700200004	119
10	2017-01-02	700200004	165
11	2017-01-02	700200004	117
12	2017-01-02	700200005	101
13	2017-01-02	700200006	353
14	2017-01-02	700200007	74

1. Calculamos el soporte de cada ítem.

Comenzamos con la forma en que se almacena la información de las transacciones

```
transactions.head(15)
```

	Visit_date	Visit_nbr	Category
0	2017-01-02	700200000	3
1	2017-01-02	700200001	126
2	2017-01-02	700200001	130
3	2017-01-02	700200002	127
4	2017-01-02	700200002	38
5	2017-01-02	700200002	130
6	2017-01-02	700200002	15
7	2017-01-02	700200002	125
8	2017-01-02	700200003	37
9	2017-01-02	700200004	119
10	2017-01-02	700200004	165
11	2017-01-02	700200004	117
12	2017-01-02	700200005	101
13	2017-01-02	700200006	353
14	2017-01-02	700200007	74

1. Calculamos el soporte de cada ítem.
2. Generamos las combinaciones de pares posibles por cada 'Visit_nbr'.

Comenzamos con la forma en que se almacena la información de las transacciones

```
transactions.head(15)
```

	Visit_date	Visit_nbr	Category
0	2017-01-02	700200000	3
1	2017-01-02	700200001	126
2	2017-01-02	700200001	130
3	2017-01-02	700200002	127
4	2017-01-02	700200002	38
5	2017-01-02	700200002	130
6	2017-01-02	700200002	15
7	2017-01-02	700200002	125
8	2017-01-02	700200003	37
9	2017-01-02	700200004	119
10	2017-01-02	700200004	165
11	2017-01-02	700200004	117
12	2017-01-02	700200005	101
13	2017-01-02	700200006	353
14	2017-01-02	700200007	74

1. Calculamos el soporte de cada ítem.
2. Generamos las combinaciones de pares posibles por cada 'Visit_nbr'.
3. Calculamos el soporte conjunto de pares de ítems posibles.

Comenzamos con la forma en que se almacena la información de las transacciones

```
transactions.head(15)
```

	Visit_date	Visit_nbr	Category
0	2017-01-02	700200000	3
1	2017-01-02	700200001	126
2	2017-01-02	700200001	130
3	2017-01-02	700200002	127
4	2017-01-02	700200002	38
5	2017-01-02	700200002	130
6	2017-01-02	700200002	15
7	2017-01-02	700200002	125
8	2017-01-02	700200003	37
9	2017-01-02	700200004	119
10	2017-01-02	700200004	165
11	2017-01-02	700200004	117
12	2017-01-02	700200005	101
13	2017-01-02	700200006	353
14	2017-01-02	700200007	74

1. Calculamos el soporte de cada ítem.
2. Generamos las combinaciones de pares posibles por cada 'Visit_nbr'.
3. Calculamos el soporte conjunto de pares de ítems posibles.
4. Utilizamos esta información para calcular los indicadores propuestos.

RESULTADOS

MINI-EJEMPLOS

Veamos con algunos ejemplos, el soporte individual y conjunto

item _a	item _b	supp _a	supp _b	supp _{a,b}
Shampoo	Pasta	13757	28852	4207
Shampoo	Leche	13757	70143	6218
Pasta	Shampoo	28852	13757	4207
Pasta	Leche	28852	70143	15384
Leche	Shampoo	70143	13757	6218
Leche	Pasta	70143	28852	15384

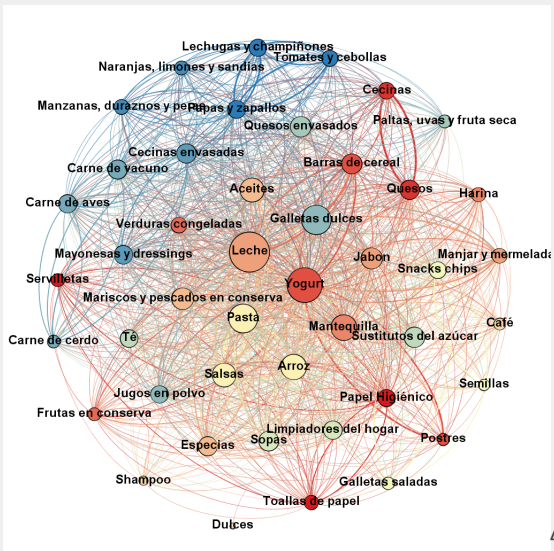
Donde hemos usado el soporte como la frecuencia de aparición de cada ítem y par.

MINI-EJEMPLOS

item _a	item _b	conf _{a→b}	Lift _{a→b}	Str _{a,b}	Conv _{a→b}
Shampoo	Pasta	0,145	23,990	0,0038	1,1635
Shampoo	Leche	0,0886	14,5852	0,0057	1,0906
Pasta	Shampoo	0,145	23,990	0,0038	1,1557
Pasta	Leche	0,5332	17,2059	0,01429	2,1149
Leche	Shampoo	0,0886	14,5852	0,0057	1,0632
Leche	Pasta	0,5332	17,2059	0,01429	2,0758

OTROS ENFOQUES

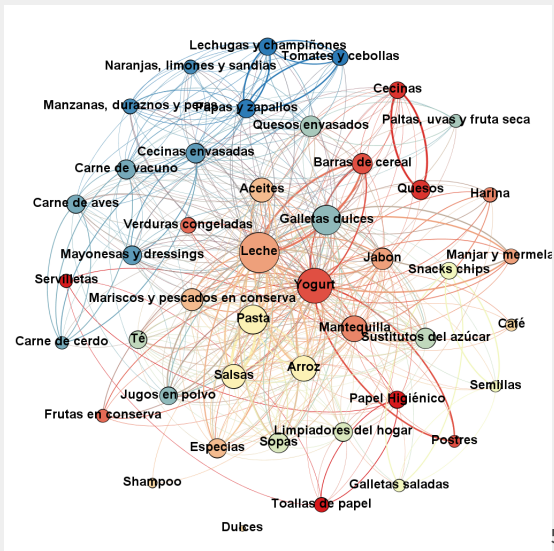
GRAFOS DIRECCIONADOS



4

$${}^4conv_{a \rightarrow b} \in [1.05, 2.85]$$

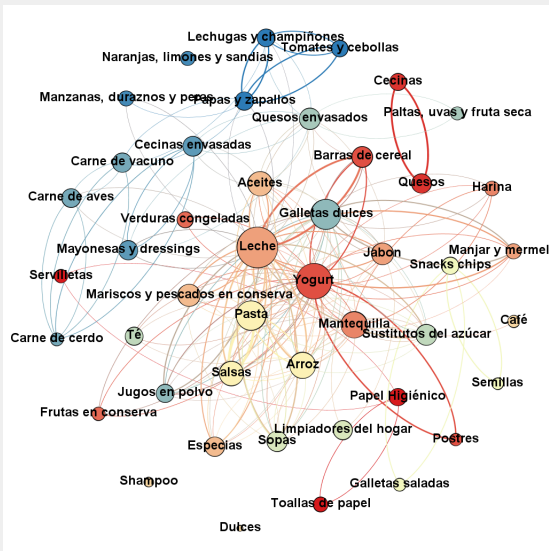
GRAFOS DIRECCIONADOS



5

$${}^5conv_{a \rightarrow b} \in [1.25, 2.85]$$

GRAFOS DIRECCIONADOS



6

$${}^6 conv_{a \rightarrow b} \in [1.3, 2.85]$$

Nuevos desarrollos e implementación de técnicas

- Incorporación de Conceptos de Topología en técnicas de análisis de Datos
- Incorporación de Conceptos de Topología Algebraica y Geometría en técnicas de análisis de Datos

⁷G. Carlsson, “Topology and Data”

UN POCO DE ANÁLISIS DE DATOS TOPOLÓGICO

Nuevos desarrollos e implementación de técnicas

- Incorporación de Conceptos de Topología en técnicas de análisis de Datos
- Incorporación de Conceptos de Topología Algebraica y Geometría en técnicas de análisis de Datos

Razones para considerar este enfoque [2]⁷

- Enfoque en información cualitativa desde propiedades generales de los datos.
- Mayor relevancia a la agregación de datos por sobre datos o parámetros específicos.
- No naturalidad de la asignación de coordenadas.

⁷G. Carlsson, “Topology and Data”

Dado que el estudio desde la homología persistente depende de la topología/geometría de la variedad ambiente de los datos [5]⁸ (o que aproxime su distribución), imponer una distribución de datos por sobre otra para el análisis de invariantes topológicos puede afectar la naturaleza de los resultados de forma artificial.

⁸L. Wasserman, “Topological Data Analysis”

Dado que el estudio desde la homología persistente depende de la topología/geometría de la variedad ambiente de los datos [5]⁸ (o que aproxime su distribución), imponer una distribución de datos por sobre otra para el análisis de invariantes topológicos puede afectar la naturaleza de los resultados de forma artificial.

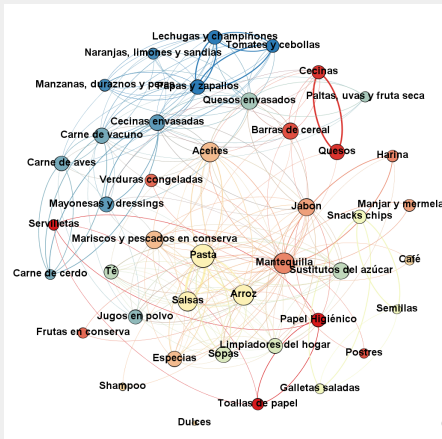
Desafíos futuros

La definición de métricas o disposiciones espaciales a partir de los datos es un problema abierto, se propone explorar posibilidades con esto en mente para trabajos futuros.

⁸L. Wasserman, “Topological Data Analysis”

POSIBILIDADES SOBRE GRAFOS DIRECCIONADOS






Veamos uno de los grafos anteriores, pero eliminando las interacciones de los items con mayor *conviction*



9

⁹ $conv_{a \rightarrow b} \in [1.25, 2.85]$ sin items con alta *conviction*.

BIBLIOGRAFIA

-  R. AGRAWAL, T. IMIELIŃSKI, AND A. SWAMI, *MINING ASSOCIATION RULES BETWEEN SETS OF ITEMS IN LARGE DATABASES*, SIGMOD REC., 22 (1993), P. 207–216.
-  G. CARLSSON, *TOPOLOGY AND DATA*, BULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY, 46 (2009), PP. 255–308.
-  M. HAHSLER, *A PROBABILISTIC COMPARISON OF COMMONLY USED INTEREST MEASURES FOR ASSOCIATION RULES*, 2005.
-  P.-N. TAN, V. KUMAR, AND J. SRIVASTAVA, *SELECTING THE RIGHT OBJECTIVE MEASURE FOR ASSOCIATION ANALYSIS*, INFORMATION SYSTEMS, 29 (2004), PP. 293–313.
KNOWLEDGE DISCOVERY AND DATA MINING (KDD 2002).
-  L. WASSERMAN, *TOPOLOGICAL DATA ANALYSIS*, ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION, 5 (2018), PP. 501–532.