Minería de patrones frecuentes y reglas de asociación

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y análitica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Métricas de rendimiento y evaluación de reglas de asociación

Soporte y confianza como medidas básicas

UCO ONLINE

- La minería de reglas de asociación se debe enfocar en generar reglas de asociación simples
- La **longitud de una regla** puede ser limitada por un *threshold* (umbral) definido por el científico de datos
  - Con un menor número de *itemsets*, la interpretación de las reglas resulta más intuitiva
  - Sin embargo, simplificar las reglas puede aumentar notablemente su número
- Los valores cuantitativos pueden agruparse y categorizarse (p.ej. En grupos de edad)

- Los algoritmos de reglas de asociación tienden a producir una gran cantidad de reglas

  muchas de ellas **no son interesantes** o **son redundantes**

  redundantes si $\{A,B,C\} \rightarrow \{D\}$ y $\{A,B\} \rightarrow \{D\}$
  tienen el mismo soporte y confianza

- Las **métricas de interés** se pueden usar para podar/ordenar los patrones
- En la formulación original de reglas de asociación se utilizan **soporte** y **confianza** pero las únicas métricas existentes

- **Soporte**: proporción del número de instancias que cubren el **antecedente** y el **consecuente** de la regla **sobre el total de instancias** del conjunto de datos.
  - Esta medida de calidad actuará mayoritariamente como medida de *fitness* en las propuestas evolutivas

- **Confianza**: proporción del número de instancias que cubren el **antecedente** y el **consecuente** de la regla sobre el número de instancias que cubren en **antecedente** de dicha regla

R. Agrawal, T. Imielinski, and A. Swami. *Mining associations between sets of items in large databases*. In Proc. of the ACM SIGMOD Int'l Conference on Management of Data, pages 207-216, Washington D.C., May 1993

Soporte

- La **utilidad de una regla** puede medirse con un **umbral de soporte** mínimo
- Las reglas para transacciones cuyos *itemsets* no están suficientemente contenidos en ambos lados (definido por un valor de umbral) pueden ser excluídos
- El **soporte** puede definirse como:

$$supp(X) = \frac{|\{T_k \in D \mid X \subseteq T_k\}|}{|D|}$$

Esta relación compara el número de eventos conteniendo un *itemset* **X** con el número de eventos de la base de datos **D**

Soporte

- Supongamos la siguiente base de datos:

$$D = \{(1,2,3), (2,3,4), (1,2,4), (1,2,5), (1,3,5)\}$$

- El **soporte** para el itemset (1,2) es:

$$supp((1,2)) = \frac{|\{T_k \in D \mid X \subseteq T_k\}|}{|D|} = 3/5 = 60\%$$

Soporte

El **soporte** del conjunto de elementos {***Pan, Mantequilla***} será **3/5=0,6**. Cualquier regla que se forme a partir de dicho conjunto de elementos tendrá el valor de **soporte 0,6**

	Leche	Pan	Mantequilla	Galletas
Usuario 1	✓	✓	✓	
Usuario 2		✓	✓	
Usuario 3	✓			✓
Usuario 4	✓	✓	✓	
Usuario 5		✓		✓

Confianza

- La **confianza de una regla** (o ***strength***) mide la frecuencia con la que un itemset encontrado en la parte izquierda de la regla, también se encuentra en la parte derecha
  - Este parámetro puede evaluarse a partir de un umbral mínimo de confianza
- Las reglas para eventos cuyos itemsets no están suficientemente contenidos en la parte derecha, aunque sí en la izquierda (definido por un valor umbral) pueden excluirse
- La **confianza** se define como:

$$conf(X_a, X_c) = \frac{supp(X_a \cup X_c)}{supp(X_a)}$$

Esta relación compara el número de eventos que contienen los itemsets $\mathbf{X_a}$ y $\mathbf{X_c}$ con el número de eventos que sólo contiene $\mathbf{X_a}$

donde existe una regla $X_a \rightarrow X_c$, de modo que los *itemsets* X_a y X_c son subregiones de un evento T_k, esto es: $X_a \subseteq T_k \wedge X_c \subseteq T_k$

Además, se cumple que $X_a \cap X_c = \emptyset$

Confianza

- Supongamos la siguiente base de datos:

$$D = \{(1,2,3), (2,3,4), (1,2,4), (1,2,5), (1,3,5)\}$$

- Calculamos la confianza de la regla 1 -> 2

$$conf((1,2)) = \frac{supp(1 \cup 2)}{supp(1)} = \frac{3/5}{4/5} = ¾ = 75\%$$

Esto es, la relación de transacciones que contienen X_a y X_c por las transacciones que contienen el itemset X_a

Confianza

A partir del conjunto de elementos {***Pan, Mantequilla***} pueden obtenerse las reglas:

- Pan ⟶ Mantequilla con una confianza de 3/4 = 0,75
- Mantequilla ⟶ Pan con una confianza de 3/3 = 1

	Leche	Pan	Mantequilla	Galletas
Usuario 1	✓	✓	✓	
Usuario 2		✓	✓	
Usuario 3	✓			✓
Usuario 4	✓	✓	✓	
Usuario 5		✓		✓

- Si la confianza alcanza un valor del 100 %, entonces se dice que la implicación es una **regla exacta**
- Aunque la confianza alcance valores altos, **la regla no será útil** a menos que el soporte también alcance valores altos
- A las reglas que tienen alta confianza y alto suporte se les denomina **reglas fuertes**

  - ¿Cómo optimizar soporte y confianza?
  - ¿Hay casos en los que interese soporte o confianza muy bajo?
    - Algunas propuestas competitivas (distintas a Apriori) pueden generar reglas útiles aún incluso con valores bajos de soporte

Métricas de rendimiento y evaluación de reglas de asociación

Otras medidas de evaluación

UCO ONLINE

S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. *Dynamic itemset counting and implication rules for market basket data*. In Proc. of the ACM SIGMOD Intl Conf. on Management of Data (ACM SIGMOD'97), pp 265-276, 1997.

Lift

- **Lift**: El **interés** o medida de **lift**, calcula cuántas veces ocurre el antecedente y el consecuente más de lo esperado en un *dataset* suponiendo que tanto el antecedente como el consecuente **son independientes**

$$\mathrm{lift}(R_{A\rightarrow C}) = \frac{\mathrm{conf}(R)}{\mathrm{supp}(C)} = \frac{\mathrm{supp}(R)}{\mathrm{supp}(A) * \mathrm{supp}(C)}$$

- Puesto que el denominador es el producto de soportes del antecedente y consecuente, el **lift** es una **medida simétrica**

lift (*Pan* → *Mantequilla***)** = **lift (***Mantequilla* → *Pan***)**

Lift

- Soporte (Pan) = 4/5 = 0,8
- Soporte (Mantequilla) = 3/5 = 0,6
- Soporte (Mantequilla) * Soporte (Pan) = 0,8*0,6 = 0,48
- Lift (Pan → Mantequilla) = 0,6/0,48 = 1,25

	Leche	Pan	Mantequilla	Galletas
Usuario 1	✓	✓	✓	
Usuario 2		✓	✓	
Usuario 3	✓			✓
Usuario 4	✓	✓	✓	
Usuario 5		✓		✓

Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules. Knowledge Discovery in Databases, 1991: p. 229-248.

UCO ONLINE

Leverage

- ***Leverage***: Calcula la diferencia entre el número de veces que ocurre el antecedente y el consecuente en un dataset y lo que se esperaba suponiendo ambos **son independientes**
  - Presenta cierta similitud con el interés (*lift*).

$$\text{leverage}(R_{A \rightarrow C}) = \text{supp}(R) - [\text{supp}(A) * \text{supp}(C)]$$

- Como la resta es con el producto de los soportes del antecedente y consecuente, el ***leverage*** es una **medida simétrica**

leverage (*Pan* → *Mantequilla***)** = **leverage (***Mantequilla* → *Pan***)**

Leverage

- Soporte (Pan) = 4/5 = 0,8
- Soporte (Mantequilla) = 3/5 = 0,6
- Soporte (Mantequilla) * Soporte (Pan) = 0,8*0,6 = 0,48
- Leverage (Pan → Mantequilla) = 0,6-0,48 = 0,12

	Leche	Pan	Mantequilla	Galletas
Usuario 1	✓	✓	✓	
Usuario 2		✓	✓	
Usuario 3	✓			✓
Usuario 4	✓	✓	✓	
Usuario 5		✓		✓

All-confidence

- **All-confidence**: representa que todas las reglas que pueden ser generadas a partir del itemset **Z** tienen al menos la confianza de **all-confidence(Z)**.
  - Posee la propiedad de clausura hacia abajo por menor conjunto (*downward-closed closure property*)

$$\text{all-confidence}(Z) = \frac{\text{supp}(Z)}{\max(\text{supp}(z \text{ de } Z))}$$

Una colección C de conjuntos es cerrada hacia abajo (**downward closed)** si para cualquier X conjunto de la colección, entonces cualquier subconjunto de X también pertenece a C

donde max(..) es el soporte del ítem con el mayor soporte en R.

All-confidence

- All-confidence (Usuario1) = supp (Usuario1) / max (0.6, **0.8**, 0.6) = (2/5) / 0,8 = 0,5
- Soporte (Leche) = 3/5 = 0.6
- Soporte (Pan) = 4/5 = 0.8
- Soporte (Leche) = 3/5 = 0.6

	Leche	Pan	Mantequilla	Galletas
Usuario 1	✓	✓	✓	
Usuario 2		✓	✓	
Usuario 3	✓			✓
Usuario 4	✓	✓	✓	
Usuario 5		✓		✓

Cobertura

- **Coverage**: representa las veces que una regla puede ser aplicada en la base de datos
- También se le conoce como soporte del antecedente

$$coverage(R_{A \rightarrow C}) = supp\ (A)$$

Cobertura

- Coverage(Leche → Galletas) = supp (Leche) = 3/5 = 0,6
- Coverage(Galletas → Leche) = supp (Galletas) = 2/5 = 0,4

	Leche	Pan	Mantequilla	Galletas
Usuario 1	✓	✓	✓	
Usuario 2		✓	✓	
Usuario 3	✓			✓
Usuario 4	✓	✓	✓	
Usuario 5		✓		✓

Evaluación de la calidad de las reglas

Dada la regla X $\rightarrow$ Y, la información necesitada se computa a partir de la **tabla de contingencia**

	Y	$\bar{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\bar{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	

f_{11}: soporte de X e Y
f_{10}: soporte de X e $\bar{Y}$
f_{01}: soporte de $\bar{X}$ e Y
f_{00}: soporte de $\bar{X}$ e $\bar{Y}$

La table permite definir varias métricas

soporte, confianza, lift, Gini,
J-measure, etc.

Evaluación de la calidad de las reglas

	Café	$\overline{\text{Café}}$	
Té	15	5	20
$\overline{\text{Té}}$	75	5	80
	90	10	100

Regla de asociación: **Té → Café**

Confianza= P(Café|Té) = 0.75

pero P(Café) = 0.9

⇒ Aunque la confianza es alta, la regla es confusa

Evaluación de la calidad de las reglas

Población de 1000 estudiantes

- 600 estudiantes saben como nadar (S)
- 700 estudiantes saben como montar en bici (B)
- 420 estudiantes saben como nadar y montar en bici ($S \wedge B$)

- $P(S \wedge B) = 420/1000 = 0.42$
- $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$

- $P(S \wedge B) = P(S) \times P(B)$ => Independencia estadística
- $P(S \wedge B) > P(S) \times P(B)$ => Positivamente correlado
- $P(S \wedge B) < P(S) \times P(B)$ => Negativamente correlado

Evaluación de la calidad de las reglas

	Café	Café	
Té	15	5	20
Té	75	5	80
	90	10	100

Regla de asociación: **Té → Café**

Confianza = P(Café|Té) = 15 / 20 = 0.75

pero P(Café) = 90 / 100 = 0.9

⇒ Lift = P(Café|Té) / P(Café) = 0.75/0.9 = 0.8333
(< 1, por lo tanto, están negativamente correladas)

Evaluación de la calidad de las reglas

	Y	Y	
X	10	0	10
X	0	90	90
	10	90	100

	Y	Y	
X	90	0	90
X	0	10	10
	90	10	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Independencia estadística:

Si P(X,Y)=P(X)P(Y) => Lift = 1

Multitud de medidas:

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\mid A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(A)})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\mid A)^2+P(\overline{B}\mid A)^2]+P(\overline{A})[P(B\mid\overline{A})^2+P(\overline{B}\mid\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\mid B)^2+P(\overline{A}\mid B)^2]+P(\overline{B})[P(A\mid\overline{B})^2+P(\overline{A}\mid\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\mid A),P(A\mid B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\mid A)-P(B)}{1-P(B)},\frac{P(A\mid B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\mid A)-P(B),P(A\mid B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\mid A)-P(B),P(A\mid B)-P(A))$

¡Gracias!

UCO ONLINE