Reglas de asociación

Mario Pérez Esteso

Introducción

Son utilizadas para encontrar reglas que describan una cierta tendencia en los datos.

Itemset

transaction ID	items
1	milk, bread
2	bread, butter
3	beer
4	milk, bread, butter
5	bread, butter

Matriz de transacciones

		items			
		i_1 milk	i_2 bread	i_3 butter	i_4 beer
itemsets	X_1	1	1	0	0
	X_2	0	1	0	1
	X_3	1	1	1	0
	X_4	0	0	1	0

Conceptos clave

Transacción: *{milk, bread} -> {butter}*

Support: supp(X -> Y)

Fracción de las transacciones que contienen a X en la parte izquierda.

Conficence: conf(X -> Y)

Relación entre el support de la regla completa y el support de la parte izquierda.

Lift.

Titanic: aprendiendo del desastre

Objetivo:

Predecir qué pasajeros sobrevivieron o no.

Datos disponibles:

Se pueden obtener del siguiente enlace, descargando el archivo `titanic.raw.rdata`: http://www.rdatamining.com/data

Análisis de los datos

¿Qué nos aportan?

- Clase en la que viaja el pasajero.
- Sexo.
- Edad.
- Sobrevive.

titanic.raw

	Class	Sex	Age	Survived
1	3rd	Male	Child	No
2	3rd	Male	Child	No
3	3rd	Male	Child	No
4	3rd	Male	Child	No
5	3rd	Male	Child	No
6	3rd	Male	Child	No
7	3rd	Male	Child	No
8	3rd	Male	Child	No
9	3rd	Male	Child	No
10	3rd	Male	Child	No
11	3rd	Male	Child	No
12	3rd	Male	Child	No

Análisis de los datos

train

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6	S
12	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103	S
13	13	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.0500		S
14	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.2750		S
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542		S
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16.0000		S
17	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250		Q
18	18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0000		S
19	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.00	1	0	345763	18.0000		S
20	20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2250		C
21	21	0	2	Fynney, Mr. Joseph J	male	35.00	0	0	239865	26.0000		S
22	22	1	2	Beesley, Mr. Lawrence	male	34.00	0	0	248698	13.0000	D56	S
23	23	1	3	McGowan, Miss. Anna "Annie"	female	15.00	0	0	330923	8.0292		Q
24	24	1	1	Sloper, Mr. William Thompson	male	28.00	0	0	113788	35.5000	A6	S
25	25	0	3	Palsson, Miss. Torborg Danira	female	8.00	3	1	349909	21.0750		S
26	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38.00	1	5	347077	31.3875		S
27	27	0	3	Emir, Mr. Farred Chehab	male	NA	0	0	2631	7.2250		C
28	28	0	1	Fortune, Mr. Charles Alexander	male	19.00	3	2	19950	263.0000	C23 C25 C27	S
29	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NA	0	0	330959	7.8792		Q
30	30	0	3	Todoroff, Mr. Lalio	male	NA	0	0	349216	7.8958		S
31	31	0	1	Uruchurtu, Don. Manuel E	male	40.00	0	0	PC 17601	27.7208		C
32	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NA	1	0	PC 17569	146.5208	B78	C
33	33	1	3	Glynn, Miss. Mary Agatha	female	NA	0	0	335677	7.7500		Q

Configuración del entorno

1. Establecer el directorio de trabajo e importar los datos:

```
setwd("~/Taller-Big-Data/Ejercicios/Titanic")
load("titanic.raw.rdata")
```

2. Importar las librerías necesarias:

```
library(Matrix)
library(arules) # install.packages("arules")
```

Creación de las reglas

3. Inspección de los datos:

```
head(titanic.raw)
```

4. Creación de las reglas de asociación:

```
rules = apriori(titanic.raw)
```

5. Inspección de las reglas:

```
inspect(rules)
```

Creación de las reglas

	lhs		rhs	support	confidence	lift
1	{}	=>	{Age=Adult}	0.9504771	0.9504771	1.0000000
2	{Class=2nd}	=>	{Age=Adult}	0.1185825	0.9157895	0.9635051
3	{Class=1st}	=>	{Age=Adult}	0.1449341	0.9815385	1.0326798
4	{Sex=Female}	=>	{Age=Adult}	0.1930940	0.9042553	0.9513700
5	{Class=3rd}	=>	{Age=Adult}	0.2848705	0.8881020	0.9343750
6	{Survived=Yes}	=>	{Age=Adult}	0.2971377	0.9198312	0.9677574
7	{Class=Crew}	=>	{Sex=Male}	0.3916402	0.9740113	1.2384742
8	{Class=Crew}	=>	{Age=Adult}	0.4020900	1.0000000	1.0521033
9	{Survived=No}	=>	{Sex=Male}	0.6197183	0.9154362	1.1639949
10	{Survived=No}	=>	{Age=Adult}	0.6533394	0.9651007	1.0153856

Creación de las reglas

6. Creación de reglas con parámetros específicos:

```
rules <- apriori(titanic.raw,
         parameter = list(minlen=1, supp=0.005, conf=0.8),
         appearance = list(rhs=c("Survived=No", "Survived=Yes"),
         default="lhs"),
         control = list(verbose=F))
```

7. Ordenar reglas con *lift* de mayor a menor:

```
rules.sorted <- sort(rules, by="lift")
```

Creación de las reglas

```
    lhs                    rhs                    support   confidence     lift
1   {Class=2nd,
    Age=Child}   => {Survived=Yes} 0.010904134  1.0000000 3.095640
2   {Class=2nd,
    Sex=Female,
    Age=Child}   => {Survived=Yes} 0.005906406  1.0000000 3.095640
3   {Class=1st,
    Sex=Female} => {Survived=Yes} 0.064061790  0.9724138 3.010243
4   {Class=1st,
    Sex=Female,
    Age=Adult}   => {Survived=Yes} 0.063607451  0.9722222 3.009650
```

Encontrar reglas redundantes

8. Crear una matriz que diga si una regla contiene a otra:

```
subset.matrix <- is.subset(rules.sorted, rules.sorted)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
redundant <- colSums(subset.matrix, na.rm=T) >= 1
```

9. ¿Cuáles son las redundantes?:

```
which(redundant)
```

Reglas redundantes

10. Eliminar reglas redundantes:

```
rules.pruned <- rules.sorted[!redundant]
```

11. Inspección de las reglas:

```
inspect(rules.pruned)
```

Visualización de reglas

12. Instalar el paquete de visualización:

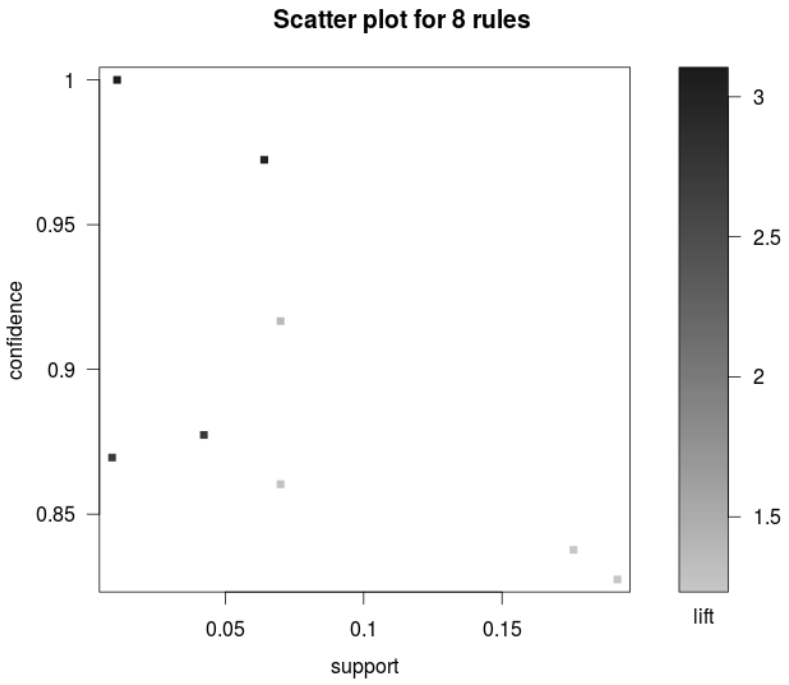
```
install.packages("arulesViz")
```

13. Importar *arulesViz*:

```
require(arulesViz)
```

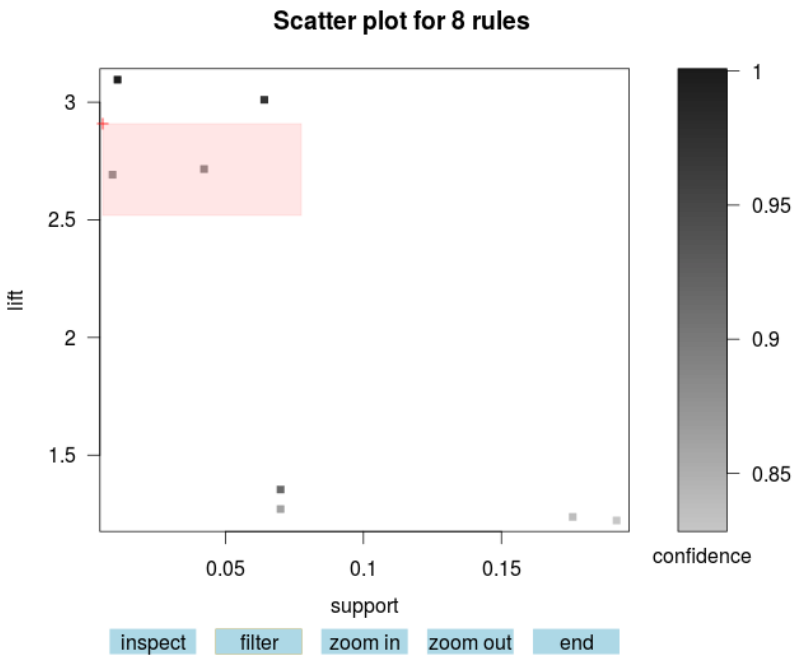
Visualización de reglas - Scatter plot

```
plot(rules.pruned)
```



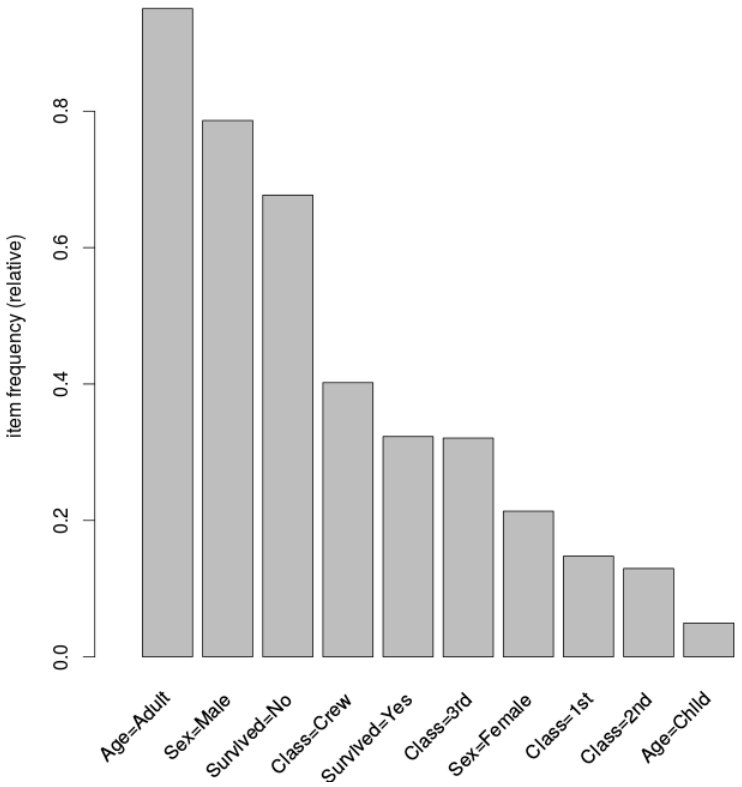
Visualización de reglas - Scatter plot

```
plot(rules.pruned,
  measure=c("support", "lift"),
  shading="confidence",
  interactive=TRUE)
```



Visualización de reglas - Histograma

```
transactions <- as(titanic.raw,
                   "transactions")
itemFrequencyPlot(transactions,
                   topN=20,
                   type="relative")
```



Visualización de reglas - Histograma

Filtrar por categoría:

```
transactions <- as(titanic.raw["Survived"],
                   "transactions")
itemFrequencyPlot(transactions,
                  type="relative")
```