ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Laboratorio de: Inteligencia de negocios
Práctica No.: 5 de Datamining
Tema: Reglas de asociación

Nombre: Díaz Padilla Danny Sebastián **Fecha:** 27/01/2020

1. Objetivos:

1.1. Objetivo General

Crear reglas de asociación para un conjunto de datos.

1.2. Objetivos Específicos

Usar weather nominal para determinar reglas de asociación usando WEKA.
De forma analítica observar un conjunto de datos y determinar los productos más frecuentes en orden de obtener reglas de asociación.

2. Marco teórico:

WEKA

Es un software programado en Java que está orientado a la extracción de conocimientos desde bases de datos con grandes cantidades de información.[1]

Reglas de asociación

Estudian las ocurrencias de un evento dentro de un conjunto de datos, dadas las diferentes medidas de interés (ítems), por ello es importante conocer la estructura en que viene la base de datos, puesto que el método de aprendizaje que presentan las reglas deben ser coherente con el objeto de estudio.

Ventajas de la regla de asociación

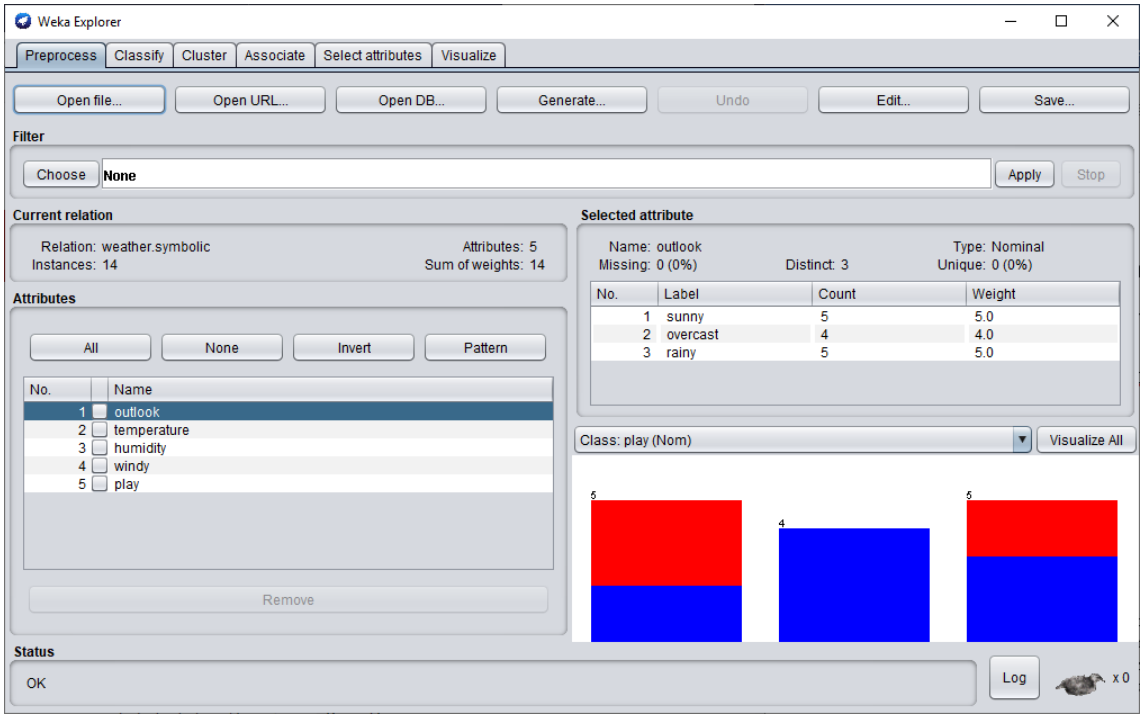
- Las asociaciones pueden darse entre cualquiera de los atributos
- Al tener muchas reglas se pueden obtener muchas conclusiones
- Al estar sujetas a un número de restricciones (acá radica la importancia en la estructura de la base de datos), la comprobación y fiabilidad es alta.

Desventajas. - El procesamiento de dicho algoritmo puede y tiende ser más lento que los de radom forest. [2]

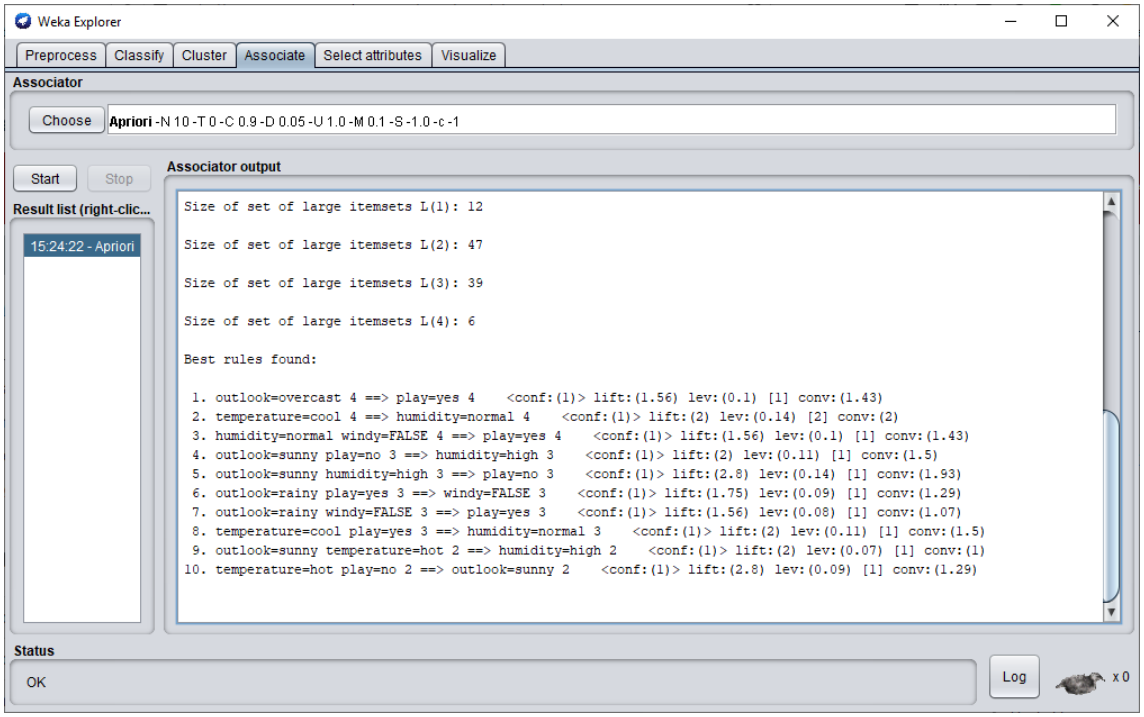
3. Desarrollo de la práctica:

A) Utilizando el archivo weather nominal arff y weka encontrar:

Exploración de los datos



Generación de las asociaciones



- Los niveles de confianza y soporte iniciales

```
Apriori
=======

Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
```

- Número de itemsets de longitud 1, 2, 3 y 4

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 6
```

Reglas

```
Best rules found:

 1. outlook=overcast 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
 2. temperature=cool 4 ==> humidity=normal 4    <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
 3. humidity=normal windy=FALSE 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
 4. outlook=sunny play=no 3 ==> humidity=high 3    <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
 5. outlook=sunny humidity=high 3 ==> play=no 3    <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)
 7. outlook=rainy windy=FALSE 3 ==> play=yes 3    <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)
 8. temperature=cool play=yes 3 ==> humidity=normal 3    <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
 9. outlook=sunny temperature=hot 2 ==> humidity=high 2    <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)
```

Aplicar el algoritmo Apriori a los datos de la figura 1 con minsup del 75%

transid	*custid*	*date*	*item*	*qty*
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4

El primer paso es contar el número de ocurrencias, llamado la confianza, de cada ítem separadamente, escaneando la base una primera vez. Se obtienen los siguientes resultados

Cantidad de items	4

Ítem	Conteo	Soporte
Pen	4	100%
Ink	3	75%
Milk	3	75%
Juice	2	50%

El próximo paso es generar la lista de todos los pares de ítems frecuentes.

Item	Conteo	Soporte
{Pen, Ink}	3	75%
{Pen, Milk}	3	75%
{Pen, Juice}	2	50%
{Ink, Milk}	2	50%
{Ink, Juice}	2	50%
{Milk, Juice}	1	25%

4. Análisis de resultados:

Excluyendo los grupos que contengan a los elementos no frecuentes, se obtiene que no existen conjuntos de tres elementos frecuentes y que:

{Pen, Ink}
{Pen, Milk}

Son dos conjuntos de datos frecuentes y visto de forma individual: Pen, Ink, Milk son frecuentes.

5. Conclusiones y recomendaciones:

- Se logró crear un conjunto de reglas de asociación por medio de WEKA en el archivo weather_nominal.arff
- Con Excel se realizó el análisis de las frecuencias (Algoritmo A priori) para determinar los elementos más frecuentes.
- Solo es necesaria una tabla de transacciones para generar reglas de asociación.

- Se recomienda personalizar los datos de configuración de las reglas de asociación para obtener un resultado más certero.
- Es recomendable observar un conjunto grande de datos para determinar correctamente la frecuencia de la población y no sesgarnos al analizar una muestra

6. Bibliografía:

[1]"Introducción a la minería de datos con Weka", *Locualo.net*, 2007. [Online]. Available: http://www.locualo.net/programacion/introduccion-mineria-datos-weka/00000018.aspx. [Accessed: 21- Jan- 2020].

[2]"Reglas de asociación", *Jimenez Carlos, 2018*. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/367334_353f1bbf1b3543e180bb9210e711a73f.html [Accessed: 21- Jan- 2020].