UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 3: Reglas de asociación

Integrantes:	Nicolás López
	Alberto Rodríguez
Curso:	Análisis de Datos
Sección	0-A-1
Profesor:	Max Chacón
Ayudante:	Javier Arredondo

8 de Enero de 2021

Tabla de contenidos

1. Introducción

Algunos de los problemas mas frecuentes que tienen los gerentes de de supermercados donde existen una gran cantidad de productos, son que productos colocar en venta, como diseñar los cupones de ventas y como colocar la mercadería en los estantes para maximizar las ventas.

Una de las herramientas para solucionar estos problemas son las reglas de asociación que es muy popular para la minería de datos en bases de datos. Su función principal es encontrar valores conjuntos para una serie de características, es decir el problema nace de desde el problema de encontrar patrones en bases de datos comerciales.

En el presente laboratorio se busca trabajar una base de datos ya conocida de hongos de la familia Agaricus y Lepiota. Este data set fue extraído de la guía de campo de la sociedad Aubudon. Todos los atributos del data set corresponden a variables categóricas.

El objetivo general de esta experiencia es realizar minería de reglas al set de datos anteriormente mencionado para así obtener reglas de asociación y usarlas como clasificadores, todo esto usando el lenguaje de programación R.

Objetivos específicos a cumplir son ver si se debe realizar la eliminación de algunos atributos en el transcurso del experimento, poder realizar un análisis adecuado a las reglas de asociación obtenidas, junto con realizar una comparación adecuada con el trabajo de Wlodzislaw explicado en los laboratorios anteriores.

A lo largo de documento se entrega un marco teórico para introducir en la materia al lector y luego realizar el análisis y manejo de la base de datos de hongos. Por último, se entrega una conclusión a partir de los resultados que se obtienen.

2. Marco Teórico

2.1. Reglas de asociación

Las reglas de asociación son aplicadas normalmente en la minería de datos para encontrar patrones y comportamientos de objetos que suelen ocurrir en conjunto. Explicado de otra forma, dichas reglas son una implicación, que se puede definir de la siguiente forma:

$$X \Rightarrow Y$$

Donde X e Y son diferentes conjuntos. Esto quiere decir, que si existe dicho comportamiento en X, este también se reflejará en Y con un cierto nivel de confianza (expresado comúnmente en porcentaje) entregado por el estudio. (Martínez, 2020)

2.2. Medidas de calidad

A partir de una o más reglas de asociación dadas, está puede ser medida con distintos parámetros:

- Confianza: Es la probabilidad condicional de que un ítem que se encuentra en un conjunto X este también esté presente en su conjunto Y de implicación. Se puede representar de la siguiente forma:

$$soporte(X \Rightarrow Y) = \frac{soporte(X \bigcup Y)}{soporte(X)}$$

- Soporte: Es la probabilidad de que el conjunto X e Y de la implicación se encuentre en el mismo conjunto entero de datos. Se puede representar de la siguiente forma:

$$conf(X \Rightarrow Y) = soporte(X \bigcup Y)$$

- Lift: Cuantifica la relación entre X e Y. Se puede expresar matemáticamente y dependiendo del resultado se puede deducir el significado de esta unidad de medida.

$$lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{soporte(Y)}$$

Si el resultado es mayor a 1 significa que es propable encontrar al conjunto X en Y. Por otro lado, si el resultado es 1 se puede decir que los conjuntos son independientes. Por último, se puede decir que existe una correlación negativa, es decir, que no se encontrará el conjunto X en Y.

2.3. Monotonicidad

El principio de monotonicidad es utilizado para descartar ítem que no son necesarios en las reglas asociadas. Asimismo, dichas reglas generadas se reducen. Esto se realiza de tal forma que si un objeto es frecuente, entonces los elementos de su conjunto potencia también lo son. En el caso contrario, si no son frecuentes los objetos, los conjuntos que contengan dicho objeto, tampoco será frecuente.

2.4. Algoritmo apriori

El algoritmo apriori es aquel que permite generar reglas de asociación a partir de itemsets frecuentes. Es un algoritmo que se puede representar como un árbol que presente los subconjuntos de un conjunto, y dice que a partir de un conjunto raíz frecuente se puede concluir que las ramas que lo componen también son frecuentes. (Martínez, 2020)

A partir del algoritmo apriori es posible obtener reglas e itemsets frecuentes. Además, dependiendo del problema, se puede limitar o no la cantidad de elementos frecuentes que se identifiquen en el estudio.

3. Obtención de Reglas

3.1. Pre-procesamineto

Antes de obtener las reglas, primero se buscó borrar variables o cambiar algún tipo de dato. En un principio los datos provienen del dataset de hongos, vienen con nombres poco carasterísticos ya que solo utilizan las iniciales, por esto, la primera modificación a la base de datos es otorgar nombres que identifiquen de mejor forma los atributos.

Luego de tener los atributos ordenados, se analiza la variables que deben ser eliminadas de la base de datos. A simple vista, solo se elimina un tipo de datos. El atributo eliminado es "veil-type" ya que solo puede tomar un valor (la cual es "partial"). No fue necesario ningún procesamiento de discretización ya que todos los atributos del set de datos con el que se trabaja son discretos.

3.2. Obtención inicial de reglas

Con los datos listos para trabajar, se inicia la obtención de reglas. Para encontrar las posibles reglas se utiliza el algoritmo apriori que genera las reglas de asociación a partir de *itemsets* frecuentes. Utilizamos la función nativa de R, *apriori*. (Maechler, 2019).

Inicialmente se iba a trabajar con un soporte mínimo de 20 %, junto con un mínimo de 1 antecedente y un máximo de 4 antecedentes, pero se crean 3781 reglas, lo cual al ser demasiadas no es una buena forma para predecir los comportamiento de los datos. Así que se intenta con un soporte mínimo mayor.

Se aumentó con un soporte mínimo de 30 %, lo cual generaba 106 reglas. Aún al ser un número de reglas muy grandes, se trabajó finalmente con un soporte mínimo de 35 %, con el mismo mínimo y máximo de antecedentes. Se olvidó mencionar que lo que importa es determinar si un hongo es venenoso o comestibles, al algoritmo se le indica específicamente que este fuera el consecuente a buscar.

El conjunto de reglas obtenido se ordenó de dos formas, por la confianza y por el soporte. Se obtuvieron 23 reglas, pero solo se escogieron las 10 primeras. De estas nos enfocamos en las reglas ordenadas respecto al soporte, ya que entre las 10 primeras reglas habían 2

reglas que tenían como consecuente que la clase "venenoso" (la regla 7 y 8 específicamente). Se observaron detalladamente cada uno de los atributos de estas reglas y se descubrió la existencia de 3 atributos que se repetían tanto en las reglas para obtener comestibles como venenosos. Uno de ellos es "gill.attachment" (branquias) como se muestra en el Cuadro 1 el valor "attached" se encuentra en 7913 de las 8123 observaciones. También se encontró el atributo "veil_color" (color de velo), en el cual como lo muestra el Cuadro 2, el valor "white" se encuentra en 7923 observaciones. Y por último el atributo "ring_number" (número de anillos) que como muestra el Cuadro 3, el valor "one" se encuentra en 7487 observaciones, para este atributo somos un pocos mas estrictos ya que el valor "two" se encuentra en 600 observaciones, lo cual no es un número tan pequeño comparado con los atributos anteriores en cual un valor superaba por demasiado a las demás. A pesar de lo mencionado anteriormente igual se escogió.

Gill attachment	**Valores**
attached	210
free	7913

Cuadro 1: Valores del atributo branquias

Veil color	**Valores**
brown	96
orange	96
white	7923
yellow	8

Cuadro 2: Valores del atributo color de velo

Con lo mencionado, se eliminaron esos 3 atributos, ya que no aportan información al momento de generar reglas lógicas.

Ring number	Valores
none	36
one	7487
two	600

Cuadro 3: Valores del atributo número de anillos

3.3. Obtención final de reglas

Después de eliminar los 3 atributos explicados del conjunto de datos, volvemos a generar las reglas lógicas al data set con las mismas condiciones. Al eliminar esos atributos, las reglas bajaron de 23 a 9 reglas lógicas. El Cuadro 4 muestra las reglas ordenadas decrecientes respecto a la confianza.

N°	Antecedentes	Consecuentes	Confianza	Soporte	Lift
1	odor=none $\wedge gill\_size = broad$	Comestible	0.9781	0.3959	1.8881
2	odor=none	Comestible	0.9659	0.4195	1.8647
3	gill_size=broad $\wedge stalk\_surface\_above\_ring = smooth \wedge stalk\_surface\_below\_ring = smooth$	Comestible	0.9530	0.3594	1.8396
4	gill_size=broad $\wedge stalk\_surface\_above\_ring = smooth$	Comestible	0.9398	0.41561	1.8142
5	gill_size=broad $\wedge stalk\_surface\_below\_ring = smooth$	Comestible	0.9364	0.3919	1.8077
6	gill_spacing=close $\wedge gill\_size = broad \wedge stalk\_surface\_above\_ring = smooth$	Comestible	0.9295	0.3506	1.7942
7	bruises=no $\wedge gill\_spacing = close$	Venenoso	0.9005	0.3924	1.8685
8	gill_size=broad $\wedge ring\_type = pendant$	Comestible	0.8915	0.3643	1.7210
9	stalk_surface_above_ring=smooth $\wedge ring\_type = pendant$	Comestible	0.8168	0.3683	1.5767

Cuadro 4: Reglas obtenidas

Se puede observar que varias reglas donde el consecuente es “Comestible” comparten atributos en común, solo varían algunos, se intentó combinar algunas reglas, pero nada mostró mejor resultado que lo mostrado en la tabla, por lo mencionado anteriormente, que muchas reglas tenían atributos en común.

4. Análisis de los resultados y comparación

Retrocediendo al laboratorio 1, al momento de analizar los datos y estudios realizados al dataset, se mencionó la existencia de Wlodzislaw (Wlodzislaw D, 1997), quién realizó un estudio para obtener reglas a partir del comportamiento de los datos para descubrir si un hongo era comestible o no. A partir de su estudio, el obtuvo cuatro reglas, de las cuales dos son posibles comparar con el estudio realizado actualmente. A continuación se muestra una tabla con las reglas:

Reglas Duch Wlodzislaw	**Confianza**	**Reglas estudio actual**	**Confianza**
R1: odor = none	98.52 %	R2: odor = none	96.56 %
R3: odor = none y stalk-surface-below-ring=scaly	99.9 %	R3: gillsize=broad, stalk-surf aceabovering=smooth y stalksurf acebelow-ring=smooth	95.30 %

Cuadro 5: Comparación de reglas.

Es posible observar como ambos estudios se asimilan, si bien no fueron llevados a cabo de la misma forma, estos presentan porcentajes de confianza y atributos similares. A partir del estudio realizado por Wlodzislaw, se puede analizar como a partir de las reglas este fue modificandolas para obtener una regla con un 100 % de confianza. Por otro lado, el estudio actualmente realizado fue en base a todos los datos necesarios para crear las reglas y no se tomaron datos a conveniencia, sino que las reglas obtenidas fueron a partir de todos los atributos. Es por esto, que se cree que las reglas obtenidas en este laboratorio fueron más generales y completas que el estudio comparado.

Luego, a partir del estudio del presente laboratorio, es posible analizar que es más fácil reconocer si un hongo es comestible mediante las reglas obtenidas ya que este consecuente se representa en la mayor parte de la tabla.

Por otro lado, las reglas elegidas son bastante contundentes viendo la confianza que entregan con su respectivo soporte. Además, el lift es mayor a uno, lo que quiere decir que la probabilidad de que se cumpla la regla de encontrar a los atributos en el consecuente es alta.

En cuanto al laboratorio 2, este pudo mostrar como era posible crear grupos de hongos según sus características similares y el comportamiento de los atributos en la base de datos. Sin embargo, el método de Clustering quedó claro que no era simple aplicarlo al mushrooms dataset y que otro método de agrupación sería más útil. Por esto, es que se cree que las reglas de asociación son un mejor procedimiento para agrupar a los hongos según las reglas y consecuentes determinados.

5. Conclusión

En este laboratorio se realizó minería de reglas sobre el set de datos *mushrooms*.

Se analizó dicho conjunto en busca de reglas lógicas que predijeran si los hongos eran comestibles o venenosos. En la reglas se ve un claro patrón de comportamiento, mayoritariamente para hongos comestibles, en donde destaca el antecedente $odor = none$ sobre los otros. En general a partir de las 9 reglas obtenidas se puede concluir que los hongos comestibles tienen características más marcadas que los hongos venenosos. Esto se puede apreciar claramente como 8 de las 9 reglas predicen hongos de la clase comestible.

Comparado con las reglas lógicas obtenidas por Wlodzislaw (Wlodzislaw D, 1997), las reglas de asociación obtenidas no pudieron alcanzar la misma complejidad, ya que a priori solo genera reglas simples de implicación en base a conjunciones y las de Wlodizslaw utilizan además negaciones y disyunciones.

En base a la experiencia realizada, se profundizó lo visto en cátedra, es decir, todo acerca de reglas de asociación, se aprendió a utilizar el algoritmo a priori a través de su implementación en R, analizando adecuadamente las reglas obtenidas y entendiendo sus estándares de calidad como confianza, soporte y lift. También se pudo comparar este laboratorio con los anteriores, lo cual fue enfatizado en el análisis de este informe. Se pudo comparar las reglas de asociación con las reglas de lógicas ya existentes. Por lo que los objetivos específicos fueron cumplidos.

Finalmente, se concluye que el laboratorio se cumplió en su totalidad ya que se realizó adecuadamente la minería de reglas al conjunto estudiado y se pudieron usar las reglas obtenidas para la clasificación.

Bibliografía

Maechler, M. (2019). *Mining Associations with Apriori*.

Martínez, C. G. (2020). *Reglas de asociación*.

Wlodzislaw D, Adamczack R, G. K. (1997). *Extraction of crisp logical rules using constructive constrained backpropagation networks*. Proceedings of International Conference on Neural Networks (ICNN'97).