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

ANÁLISIS DE DATOS: LABORATORIO 5

ÁRBOLES DE DECISIÓN

PABLO CÁCERES LUZANTO
BENJAMÍN HERNÁNDEZ CORTÉS

Profesor:
Max Chacón
Ayudante:
Adolfo Guzmán

Santiago - Chile
2 de septiembre de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

En experiencias anteriores, se ha estudiado en profundidad la base de datos 'Zoo' creada por el investigador Richard Forsyth, pero no solamente a un nivel básico como lo es el análisis estadístico trivial, sino que también a nivel de *Clustering*, Reglas de Asociación e incluso a través de Clasificadores Bayesianos. Como es sabido, existen diversos métodos algo más complejos y que funcionan de manera distintas a los mencionados anteriormente para estudiar datos y extraer conocimiento de ellos. Es por aquella razón, que en esta experiencia se intenta extraer conocimiento y clasificar datos de acuerdo al método de los Árboles de Decisión, a través del software RStudio.

1.2 ORGANIZACIÓN DEL DOCUMENTO

El presente documento está estructurado básicamente en dos grandes capítulos. El primero de ellos corresponde a la Obtención del Árbol, en donde de acuerdo a ciertos parámetros del programa que se realiza se genera un árbol de decisión adecuado para su posterior comparación con las reglas obtenidas en la experiencia anterior. Además, en este capítulo se incluye el significado de los parámetros mencionados y el porqué de su utilización (se considera un numero máximo de 30 reglas para comparar con el árbol de decisión). Finalmente, se presenta el capítulo de Comparación de los Métodos, en donde se compara el método de los árboles de decisión con las reglas de asociación pero en términos de los resultados obtenidos, cantidad de reglas, información que entrega, y la forma de mostrar los resultados para cada uno de los métodos.

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS

Para el estudio y análisis de los datos se utiliza la base de datos *"Zoo DataSet"* contenida en el archivo *zoo.data* cuya información y documentación se encuentra detallada en el archivo *zoo.names*.

Por otro lado, para realizar el estudio y manejo de los datos en sí, se hace uso del lenguaje de programación R junto al software *RStudio* en su versión estable más reciente, e instalado en un Sistema Operativo basado en *Linux*.

CAPÍTULO 2. OBTENCIÓN DEL ÁRBOL DE DECISIÓN

2.1 PRE-PROCESAMIENTO

Para obtener el clasificador, es necesario realizar un pre-procesamiento de los datos, el cual se describe detalladamente a continuación, para luego dar paso a la obtención del clasificador bayesiano en sí.

2.1.1. Eliminación de variables

La eliminación de variables se realiza considerando que la eliminación de éstas reducirá la dimensionalidad de la base de datos y además, considerando el aporte e impacto de la variable al momento de realizar la clasificación.

Variable 'animal_name'

Esta variable es eliminada debido a que, como su nombre lo indica, solo representa el nombre de los animales presentes en la muestra, y por ende, es un dato no-numérico que no realizará ningún aporte de valor al proceso de clasificación.

2.1.2. Eliminación de instancias

También puede darse la eliminación de instancias, bajo los mismos criterios mencionados en el inicio de la sección anterior.

Instancia 'girl'

En este caso, esta instancia es eliminada dado que no resulta adecuada mantenerla dentro del conjunto de datos, debido a que en sí mismo, el conjunto de datos fue destinado y confeccionado para detallar aspectos de las distintas especies animales pertenecientes a un zoológico.

Instancia 'frog'

En el caso de la instancia *'frog'*, se opta por eliminar una instancia de ésta, ya que se poseen dos y la única diferencia que se halla entre ambas es que una resulta ser venenosa y la otra no, atributo que no se considera importante o crucial a la hora de clasificar a la instancia en un grupo u otro.

2.2 ÁRBOL DE DECISIÓN

En esta oportunidad, para obtener el Árbol de decisión, se utiliza la función C5.0, que pertenece a la biblioteca C50 y que puede recibir como parámetros las reglas, el árbol generado, la matriz de confusión, entre otros. Dentro de los parámetros que se utilizan para obtener el árbol se encuentran:

- X: Matriz de predictores con los atributos estudiantes. En este caso, se utiliza un subconjunto de entrenamiento del *DataSet* con todos los atributos presentes en el *DataSet* original a excepción del mencionado en la sección anterior: *animal_name*.
- Y: Vector con los valores del atributo experto, que en este caso corresponde al atributo *type*.
- rules: Valor booleano. En este caso se establece en Verdadero para obtener las reglas que desprenden del árbol.

Luego de realizado lo anterior, se obtiene el árbol y las respectivas reglas, que se detallan en el capítulo siguiente.

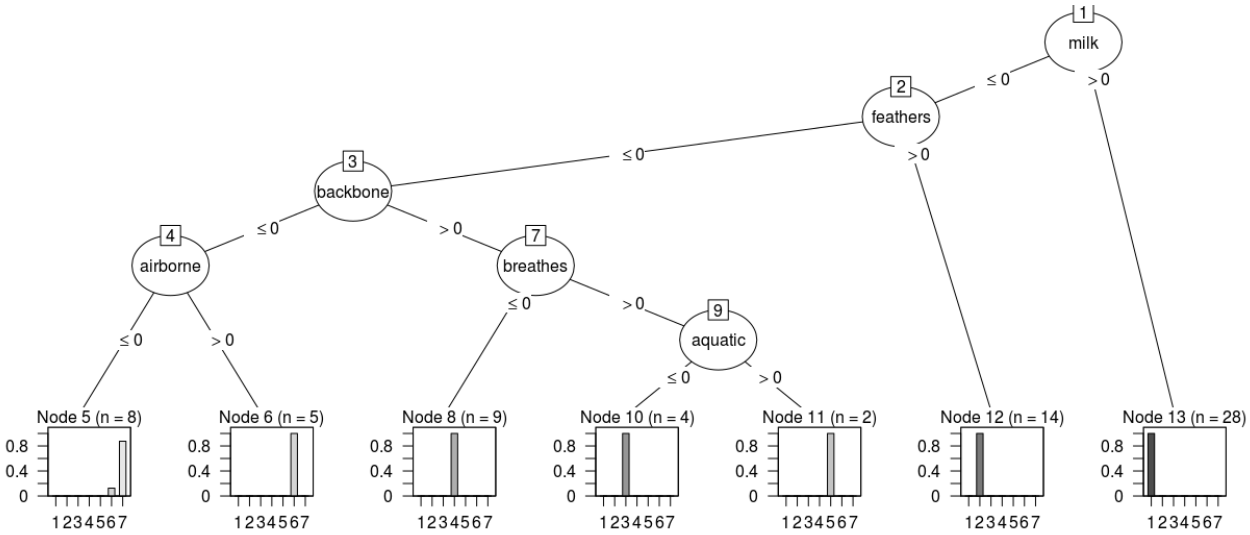

Figura 2.1: Árbol de decisión obtenido

CAPÍTULO 3. COMPARACIÓN DE LOS MÉTODOS

En primer lugar, se describe el Árbol de Decisión obtenido en esta experiencia. Como es posible apreciar en la figura 2.1, que se aprecia con mayor resolución en el Anexo del documento, es correcto señalar que cada camino desde la raíz del árbol hacia las hojas representa una regla de asociación. En este caso, entonces se tienen siete reglas, las cuales representan el tipo de animal al que pertenece cierta muestra del *DataSet* una vez pasa por las distintas bifurcaciones (antecedentes o atributos) del camino hacia la respectiva hoja (consecuente, clase, tipo de animal, etc.). Cabe destacar, que como la función utilizada para obtener el Árbol de decisión en R es la perteneciente a la librería C50, ésta emplea el algoritmo de *Quinlan* para la obtención del árbol, por lo tanto, calcula la cantidad de información que aporta cada uno de los atributos respecto al atributo experto (de manera transparente al programador), lo que básicamente se traduce en la aparición de solo algunos atributos en el árbol (los que aportan mayor información respecto a los que no aparecen). A continuación entonces, se presentan las reglas y se explica de manera detallada cada una de ellas:

1. $(milk = 0 \wedge feathers = 0 \wedge backbone = 0 \wedge airborne = 0) \rightarrow (animal.type = 7)$

   Esta regla básicamente indica que para que un animal pertenezca a la clase 'Otros'(*animal.type* = 7), éste no debe poder dar leche a sus crías (*milk* = 0), tampoco debe poseer plumas (*feathers* = 0) al igual que columna vertebral (*backbone* = 0) y además, no le sea posible volar (*airborne* = 0).

2. $(milk = 0 \wedge feathers = 0 \wedge backbone = 0 \wedge airborne = 1) \rightarrow (animal.type = 6)$

   En este caso, para que un animal pertenezca a la clase 'Insectos', es necesario que éste no pueda dar leche a sus crías (*milk* = 0), no posea plumas (*feathers* = 0), tampoco posea columna vertebral (*backbone* = 0) y, en esta oportunidad, sí tenga la posibilidad de desplazarse de manera aérea (*airborne* = 1).

3. $(milk = 0 \wedge feathers = 0 \wedge backbone = 1 \wedge breathes = 0) \rightarrow (animal.type = 4)$

   Para esta regla, que clasifica al animal como perteneciente a la clase 'Peces' (*animal.type* = 4), es necesario que este último no sea capaz de dar leche a sus crías (*milk* = 0) y tampoco posea plumas (*feathers* = 0) pero, en este caso, éste si posea columna vertebral (*backbone = 1* y no respire aire (*breathes* = 0).

4. $(milk = 0 \wedge feathers = 0 \wedge backbone = 1 \wedge breathes = 1 \wedge aquatic = 0) \rightarrow (animal.type = 3)$

   En esta regla, para que un animal sea clasificado como Reptil (*animal.type* = 3), debe no ser capaz de dar leche a sus crías (*milk* = 0), no debe poseer plumas (*feathers* = 0) y debe poseer columna vertebral (*backbone* = 1). Además, no debe vivir (en ningún momento) en el agua (*aquatic* = 0).

5. $(milk = 0 \wedge feathers = 0 \wedge backbone = 1 \wedge breathes = 1 \wedge aquatic = 1) \rightarrow (animal.type = 5)$

En este caso, para que un animal sea clasificado como Anfibio, éste debe no ser capaz de dar leche a sus crías (*milk* = 0), no debe poseer plumas (*feathers* = 0) pero sí columna vertebral (*backbone* = 1) . Además, debe poder respirar aire (*breathes* = 1) y vivir en algún momento de su vida en el agua (*aquatic* = 1).

6. $(milk = 0 \wedge feathers = 1) \rightarrow (animal.type = 2)$

   Para que un animal sea considerado perteneciente a la clase 'Aves', éste debe no ser capaz de dar leche a sus crías (*milk* = 0) y además, debe poseer plumas (*feathers* = 1).

7. $(milk = 1) \rightarrow (animal.type = 1)$

   Finalmente, para que un animal sea considerado mamífero, es necesario que éste tenga la capacidad de dar leche a sus crías (*milk* = 1).

A continuación, se exponen las reglas de asociación no-redundantes obtenidas en experiencias anteriores para su posterior análisis con el árbol de decisión obtenido en la actual experiencia. El hecho de seleccionar las no-redundantes se da por las cualidades de las mismas y que permiten asociarlas de mejor forma al árbol de decisión.

	lhs		rhs	support	confidence	lift	count
[1]	{feathers}	=>	{type=ave}	0.20202020	1	4.950000	20
[2]	{milk}	=>	{type=mamifero}	0.40404040	1	2.475000	40
[3]	{airborne,legs=6}	=>	{type=insecto}	0.06060606	1	12.375000	6
[4]	{breathes,legs=6}	=>	{type=insecto}	0.08080808	1	12.375000	8
[5]	{legs=4,domestic}	=>	{type=mamifero}	0.07070707	1	2.475000	7
[6]	{domestic,catsize}	=>	{type=mamifero}	0.05050505	1	2.475000	5
[7]	{eggs,fins}	=>	{type=pez}	0.13131313	1	7.615385	13
[8]	{airborne,aquatic}	=>	{type=ave}	0.05050505	1	4.950000	5
[9]	{eggs,legs=2}	=>	{type=ave}	0.20202020	1	4.950000	20
[10]	{hair,legs=2}	=>	{type=mamifero}	0.06060606	1	2.475000	6

Figura 3.1: Top-10 de las reglas no-redundantes ordenadas por confianza

Se dice que un regla $X \Rightarrow Y$ es redundante si existe otra regla $X' \Rightarrow Y$ tal que X' es subconjunto de X y la confianza de la última regla es mayor o igual que la primera. Matemáticamente se puede describir como:

$$\exists X' \subset X,\ conf(X' \Rightarrow Y) \geq conf(X \Rightarrow Y) \tag{3.1}$$

Con la definición anterior y en término simples, se puede entender que una regla no-redundante es aquella que posee una menor cantidad de items en el antecedente (X') en comparación a otra regla, pero que aún así permite concluir un consecuente (Y) en particular, con una confianza de igual o mayor magnitud.

A partir de la figura 2.1 y la figura 3.1, es posible observar algunas similitudes dadas para algunas reglas, específicamente para las reglas 6 y 7 detalladas previamente, las cuales se presentan en los resultados tanto del árbol de decisión como de las reglas de asociación. Sin embargo, de las demás reglas no es posible observar similitudes muy precisas o fuertes, en dónde para el árbol de decisión se concluye que un animal de tipo 'Insecto' se puede clasificar clasificar mediante 4 atributos, mientras que en las reglas de asociación esto se puede realizar identificando la presencia del atributo *airborne* o *breathes*, junto al atributo *legs=6*.

Al comparar ambos métodos, el método de árboles de decisión permite clasificar un animal en un determinado tipo mediante la presencia (o ausencia) de atributos o características que sobresalen en un mayor grado en el conjunto de datos, mientras que el método de reglas de asociación intenta encontrar reglas particulares que permitan clasificar un animal solo mediante la presencia de determinados atributos, en dónde algunos podrían resultar muy específicos, pero que ofrecen una clasificación más precisa y confiable, tal como ocurre con el atributo *legs=6* en la clasificación de 'Insectos'.

Por otra parte, es importante destacar que la cantidad de reglas generadas en la experiencia anterior es bastante superior a la cantidad de reglas generadas en la experiencia actual. Lo anterior básicamente se debe a las formas de obtención de cada una de las reglas, ya que para el Algoritmo de Quinlan (árboles de decisión) primero se identifica qué variables o atributos son los que otorgan mayor información a la 'clasificación', mientras que para el caso de las Reglas de Asociación, el algoritmo utilizado genera todas las reglas de manera independiente y posterior a esto, se filtran de acuerdo a un criterio establecido, ya sea la confianza de las reglas, el soporte o simplemente el *lift*.

CAPÍTULO 4. ANEXO