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

Laboratorio 5: Árboles de Desición

Integrantes: Nicole Reyes
Diego Águila

Curso: Análisis de Datos

Profesor: Max Chacón

Ayudante: Francisco Muñoz

24 de Junio de 2019

Tabla de contenidos

Índice de figuras

Índice de cuadros

1. Introducción

Obtener conocimiento de un conjunto de datos es fundamental al momento de realizar un estudio o análisis con respecto a estos, existen diversas técnicas y estrategias en la minería de datos para lograrlo, la obtención de un árbol de decisión es una de las técnicas conocidas para lograr esto y permite establecer un modelo predictivo el cual, muchas veces se utiliza para representar visualmente decisiones y toma de decisiones o en el caso de minería de datos, se crea un modelo que predice el valor de una variable de destino en función de diversas variables de entrada.

Esta quinta experiencia de laboratorio consiste en, mediante el proceso de minería de datos y a través de la utilización de un árbol de decisión obtener conocimiento de la base de datos asignada en la primera experiencia "*Allhypo*" correspondiente al hipotiroidismo.

1.1. Objetivos

1. Profundizar el entendimiento de los árboles de decisión como medida de obtención de conocimiento.

2. Corroborar la teoría vista en clases mediante la aplicación del árbol de decisión en el lenguaje estadístico R, haciendo uso del paquete "$C50$".

3. Obtener conocimiento de la base de datos "*Allhypo*" a través de esta técnica de minería de datos.

A continuación se define cómo se trabaja la base de datos para obtener el árbol de decisión y finalmente se realiza una comparación entre los resultados obtenidos del árbol de decisión con respecto a los resultados de las experiencias anteriores, concretamente, la experiencia número 3 correspondiente a reglas de asociación.

2. Obtención del árbol

2.1. Pre Procesamiento

Al igual que en las experiencias anteriores es fundamental elaborar una limpieza a la base de datos. En esta oportunidad la limpieza está compuesta por las siguientes etapas:

2.1.1. Eliminación de valores vacíos

Dentro de la base de datos, existen celdas con valores nulos (representados con el símbolo "?"), por lo que se opta por eliminar todos los sujetos de la muestra que contengan este símbolo para alguna variable ya que no aportan nada al análisis.

2.1.2. Eliminación de variables

Dentro del conjunto de datos usados, hay columnas que no formarán parte del estudio, debido a que no aportan información importante para este pudiendo estropear el análisis. Las variables que son eliminadas son:

- Variable TBG: Como toda su columna representa valores vacíos, no entrega información relevante debido a que se puede inferir a que no se midió la presencia de esta proteína en ningún sujeto.
- Variables de verificación de medición (measured): Indican si el paciente se ha realizado alguna medición, siendo esta información no relevante. Se decide dejar fuera las variables: *TSH measured, T3 measuted, TT4 measured, T4U measured, FTI measured y TBG measured.*
- Variable de fuente: La variable *referral.source* no entrega información que sea pertinente para el análisis ya que solo indica la fuente de la cual se obtuvo los datos de sujeto en cuestión.

2.2. Categorización de los datos

Se transforman ciertos datos a variables nominales para que se tenga de esta manera una categorización de los datos que representan esas variables. Para las hormonas, se realiza lo mismo que en la experiencia pasada de *Reglas de Asociación*. Con respecto a la media de cada una, se va clasificando a la variable, es decir, si el valor de la variable es menor a la media se deja categorizada como *'f'*, caso contrario, se deja como *'t'*. Lo mismo se hace con la edad, si el valor es menor a la media se deja categorizada como *'f'* y si es mayor, se deja como *'t'*.

2.3. Árbol de decisión

Se usa la biblioteca *C50* para la obtención del árbol cual retorna un objeto de clase C5.0 que puede ser el árbol de decisión, reglas, entre otros. Para la generación del árbol, es necesario definir su raíz, el cual debe ser el atributo adecuado que dé garantía de que los resultados que se entreguen sean lo mas cercano a la realidad. Por esta razón el atributo escogido es el *state*, el cual corresponde a la clase de Hipotiroidismo que presenta el sujeto.

Para generar el árbol, lo que se necesita es la base de datos en donde se encuentran la información necesaria y el atributo que se utilizará como raíz de este árbol. Entonces lo que se genera queda mostrado en la Figura 1:

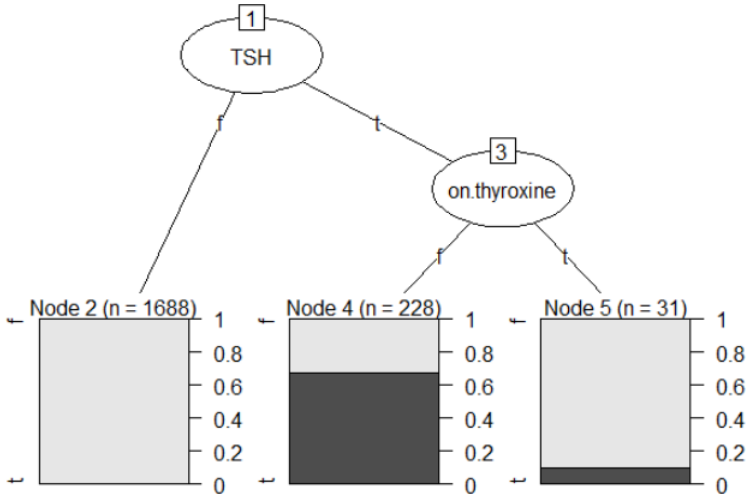

Figura 1: Árbol obtenido.

3. Análisis de los resultados

A continuación se realiza el análisis con respecto al árbol obtenido durante el desarrollo de esta experiencia. De acuerdo a lo obtenido en el modelo basado en los árboles de decisiones en el Capítulo 2, este árbol se puede expresar de acuerdo al modelo basado en *Reglas de Asociación* las cuales se encuentran descritas en el Cuadro 1 a continuación:

N° de Regla	**Antecedentes**	**Consecuente**
1	TSH=F	hypothyroid=F
2	TSH=T, on.thyroxine=F	hypothyroid=T
3	TSH=T, on.thyroxine=T	hypothyroid=F

Cuadro 1: Reglas obtenidas del árbol de decisión.

Del árbol de decisión correspondiente a la Figura 1 del Capítulo 2 y de las reglas presentadas en el Cuadro 1, se obtiene que la variable más importante resulta ser TSH, es decir, si el paciente presenta altos o bajos niveles de la hormona TSH (Tirotopina) considerándola como suficiente para determinar si un paciente no posee Hipotiroidismo, para los casos en los que la hormona TSH resulta ser alta, se incorpora la variable *on.thyroxine* considerada relevante en el árbol de decisión, la cual indica si un paciente se encuentra bajo tratamiento de reemplazo de Tiroxina, donde en los casos negativos, en su mayoría, el paciente presenta Hipotiroidismo y en los casos positivos, pocos de los pacientes presentan hipotiroidismo. Lo anterior se puede asociar a que dado el modelo de árboles de decisión de Quinlan se puede obtener la entropía de las variables que son importantes para la generación del árbol, para así saber que tan importantes son para el modelo. A continuación se muestra en el Cuadro 2, la entropía de las dos variables que se usaron para la generación del árbol:

Variable	**Entropía**
TSH	100 %
on.thyroxine	13,3 %

Cuadro 2: Entropía de las variables.

3.1. Comparación con Reglas de Asociación

En la tercera experiencia de laboratorio correspondiente a las reglas se asociación, se obtienen todas las reglas existentes en la base de datos de las cuales se obtiene un ranking con las 10 mejores a partir de medidas de calidad como el soporte, la confianza y el lift, estas reglas se presentan a continuación en el Cuadro 3:

N° de Regla	**Antecedentes**	**Consecuente**
1	on.thyroxine=F, TSH=T, TT4=F	hypothyroid=t
2	on.thyroxine=F, thyroid.surgery=F, TSH=T	hypothyroid=t
3	on.thyroxine=F, TSH=T, T3=F	hypothyroid=t
4	on.thyroxine=F, I131=F, TSH=T	hypothyroid=t
5	on.thyroxine=F, TSH=T, T4U=F	hypothyroid=t
6	on.thyroxine=F, TSH=T, FTI=F	hypothyroid=t
7	on.thyroxine=F, age=F, TSH=T	hypothyroid=t
8	on.thyroxine=F, psych=F, TSH=T	hypothyroid=t
9	on.thyroxine=F, on.antithyroid.med=F, TSH=T	hypothyroid=t
10	on.thyroxine=F, lithium=F, TSH=T	hypothyroid=t

Cuadro 3: Reglas obtenidas en la experiencia 3.

Como se observa en el Cuadro anterior, todas las reglas tienen como antecedente que el paciente posee altos niveles de la hormona TSH y no se encuentran bajo tratamiento de reemplazo de Tiroxina, además de otros factores, de los cuales tienen como consecuente, la presencia de Hipotiroidismo en el paciente. El hecho de que las 10 reglas posean como antecedente altos niveles de hormona TSH y que el paciente no se encuentre en tratamiento de reemplazo de Tiroxina, permite concluir que estas dos variables son gatillantes para determinar si el usuario posee hipotiroidismo, no así, las otras variables como TT4, T3, T4U, entre otras.

Estos resultados concuerdan con lo obtenido en el árbol de decisión, al observar el Cuadro 1 se tiene que si un paciente presenta altos niveles de hormona TSH y no se encuentra en tratamiento de reemplazo de Tiroxina tenderá a poseer Hipotiroidismo, lo cual encaja con

las 10 reglas obtenidas en la experiencia 3. Con respecto a los otros dos casos, si el paciente presenta altos niveles de hormona TSH pero se encuentra en tratamiento de reemplazo de tiroxina tenderá a no poseer Hipotiroidismo y de forma más tajante, si el paciente posee bajos niveles de hormona TSH no presentaría Hipotiroidismo.

Al obtener resultados similares a los de la experiencia 3, un árbol de decisión resulta ser una medida más eficiente para predecir la ocurrencia de un evento debido a que se construye en base a las variables que aportan una mayor cantidad de información al modelo, contrario a las reglas de asociación, donde se obtiene la totalidad de reglas a partir de una cantidad de antecedentes y de las cuales se debe realizar un filtrado posterior a modo de obtener las más relevantes en base a medidas de calidad.

4. Anexos

A continuación se presenta el script en R utilizado para el desarrollo de esta experiencia.

```
library(C50)

data <- read.csv("C:\\Users\\diego\\Dropbox\\Usach\\ANLISIS DE DATOS\\Lab\\allhypo.data",
                 col.names = c("age","sex", "on.thyroxine", "query.thyroxine", "on.antithyroid.med",
                               "sick","pregnant", "thyroid.surgery", "I131", "query.hypothyroid",
                               "query.hyperthyroid", "lithium", "goitre", "tumor", "hypopituitary",
                               "psych", "TSH.measured", "TSH", "T3.measured", "T3", "TT4.measured",
                               "TT4","T4U.measured", "T4U", "FTI.measured","FTI", "TBG.measured",
                               "TBG", "referral.source","state"), header=FALSE, sep=",", stringsAsFactors=FALSE)

# LIMPIEZA DE LA BASE DE DATOS

# Se limpia la variable de estado del paciente eliminando numeros
state <- data$state
state <- strsplit(state, ".", fixed = TRUE)
state <- lapply(state, `[[`, 1)
state <- unlist(state)
data$state <- state

# Se eliminan los pacientes que contengan algun NA
# Sex
error = data$sex == '?'
data = data[!error,]
# Age
error = data$age == '?'
```

```
data = data[!error,]
# On thyroxine
error = data$on.thyroxine == '?'
data = data[!error,]
# Query on antithyroid thyroxine
error = data$query.thyroxine == '?'
data = data[!error,]
# On antithyroid medication
error = data$on.antithyroid.med == '?'
data = data[!error,]
# Sick
error = data$sick == '?'
data = data[!error,]
# Pregnant
error = data$pregnant == '?'
data = data[!error,]
# Thyroid surgery
error = data$thyroid.surgery == '?'
data = data[!error,]
# I131 treatment
error = data$I131 == '?'
data = data[!error,]
# Query hypothyroid
error = data$query.hypothyroid == '?'
data = data[!error,]
# Query hyperthyroid
error = data$query.hyperthyroid == '?'
data = data[!error,]
# Lithium
error = data$lithium == '?'
data = data[!error,]
# Goitre
error = data$goitre == '?'
data = data[!error,]
# Tumor
error = data$tumor == '?'
```

```r
data = data[!error,]
# Hypopituitary
error = data$hypopituitary == '?'
data = data[!error,]
# Psych
error = data$psych == '?'
data = data[!error,]
# TSH
error = data$TSH == '?'
data = data[!error,]
# T3
error = data$T3 == '?'
data = data[!error,]
# TT4
error = data$TT4 == '?'
data = data[!error,]
# T4U
error = data$T4U == '?'
data = data[!error,]
# FTI
error = data$FTI == '?'
data = data[!error,]

# Se eliminan las variables que indican si se realizo la medicion de la
    hormona
data$T3.measured <- NULL
data$TSH.measured <- NULL
data$TT4.measured <- NULL
data$T4U.measured <- NULL
data$FTI.measured <- NULL
data$TBG.measured <- NULL
#Se elimina la variable TBG que no fue medida para ningun paciente
data$TBG <- NULL
#Se elimina la variable referral.source
data$referral.source <- NULL

```

```
#
    ------------------------------------------------------------------------------


# Una vez limpiada la base de datos, se transforman los datos correspondientes

# Variables continuas
data$T3 <- as.numeric(data$T3)
data$age <- as.numeric(data$age)
data$TSH <- as.numeric(data$TSH)
data$TT4 <- as.numeric(data$TT4)
data$T4U <- as.numeric(data$T4U)
data$FTI <- as.numeric(data$FTI)

# Variables binarias
data$sex <- as.factor(data$sex)
data$sick <- as.factor(data$sick)
data$I131 <- as.factor(data$I131)
data$state <- as.factor(data$state)
data$tumor <- as.factor(data$tumor)
data$psych <- as.factor(data$psych)
data$goitre <- as.factor(data$goitre)
data$lithium <- as.factor(data$lithium)
data$pregnant <- as.factor(data$pregnant)
data$on.thyroxine <- as.factor(data$on.thyroxine)
data$hypopituitary <- as.factor(data$hypopituitary)
data$query.thyroxine <- as.factor(data$query.thyroxine)
data$thyroid.surgery <- as.factor(data$thyroid.surgery)
data$query.hypothyroid <- as.factor(data$query.hypothyroid)
data$query.hyperthyroid <- as.factor(data$query.hyperthyroid)
data$on.antithyroid.med <- as.factor(data$on.antithyroid.med)

#Se obtiene la media de los datos continuos para establecerla de manear
    binaria
mediaAge <- mean(data$age)
mediaT3 <- mean(data$T3)
```

```
mediaT4U <- mean(data$T4U)
mediaTT4 <- mean(data$TT4)
mediaTSH <- mean(data$TSH)
mediaFTI <- mean(data$FTI)

#Se declaran los arreglos  con las nuevas variables binarias
newAge <- integer(length(data[[1]]))
newT3 <- integer(length(data[[1]]))
newT4U <- integer(length(data[[1]]))
newTT4 <- integer(length(data[[1]]))
newTSH <- integer(length(data[[1]]))
newFTI <- integer(length(data[[1]]))

#Se transforman los datos continuos a binarios
for(i in 1:length(data[[1]])){
  #si la edad del paciente es mayor que la media, entonces es un paciente de
    edad avanzada
   if(data$age[i] >= mediaAge){newAge[i] <- "t"}else{newAge[i] <- "f"}
  #si la cantidad de hormona presente en el paciente es mayor que la media se
    considera como que posee altos niveles de esta
   if(data$T3[i] >= mediaT3){newT3[i] <- "t"}else{newT3[i] <- "f"}
   if(data$TSH[i] >= mediaTSH){newTSH[i] <- "t"}else{newTSH[i] <- "f"}
   if(data$TT4[i] >= mediaTT4){newTT4[i] <- "t"}else{newTT4[i] <- "f"}
   if(data$T4U[i] >= mediaT4U){newT4U[i] <- "t"}else{newT4U[i] <- "f"}
   if(data$FTI[i] >= mediaFTI){newFTI[i] <- "t"}else{newFTI[i] <- "f"}
}
#Se binariza la columna state, aquellos valores falsos son lo que poseen
    negative y todos los dems son verdaderos, es decir, presentan
    hipotiroidismo
data$state <- ifelse(data$state %in% c("primary hypothyroid", "secondary
    hypothyroid", "compensated hypothyroid"), "t", "f")
#Se reemplazan las columnas correspondientes
data$state <- as.factor(data$state)
data$age <- as.factor(newAge)
data$T3 <- as.factor(newT3)
data$TT4 <- as.factor(newTT4)
```

```
161 data$T4U <- as.factor(newT4U)
162 data$TSH <- as.factor(newTSH)
163 data$FTI <- as.factor(newFTI)
164
165 #Se obtienen los datos que se utilizan para construir el arbol,
166 #es decir, todas las columnas menos state(la que define si el paciente
      presenta hipotiroidismo)
167 data_tree <- subset(data, select = -state)
168
169 #Se modela el arbol con la funcin C5.0 con los datos para el arbol y la
      variable data$state
170 model <-C5.0(data_tree, data$state)
171
172 #Se plotea el arbol
173 tree <- plot(model)
```

Bibliografía

[Bomarito] Bomarito, D. M. J. Hipotiroidismo: síntomas y tratamiento. [Online] https://www.hospitalaleman.org.ar/mujeres/hipotiroidismo-sintomas-y-tratamiento/.

[2] Caparrini, F. S. (2017). Aprendizaje inductivo: Árboles de decisión. [Online] http://www.cs.us.es/~fsancho/?e=104.

[3] Chacón, M. (2015). Árboles de decisión. [Online] http://www.udesantiagovirtual.cl/moodle2/pluginfile.php?file=%2F115779%2Fmod_resource%2Fcontent%2F0%2FCapitulo%20VII%20%C3%81rboles%20de%20Decisi%C3%B3n.pdf.

[Hurtado] Hurtado, C. Arboles de decisión (i). [Online] https://www.u-cursos.cl/ingenieria/2007/1/CC52A/1/material_docente/bajar?id_material=119128.