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

INFORME LABORATORIO 5 ANÁLISIS DE DATOS

TIC-TAC-TOE ENDGAME DATA SET

ARBOLES DE DECISIÓN

CARLOS CÁCERES
DANY RUBIANO

Profesor: Felipe Bello
Ayudantes: Bryan Guzmán
Fernanda Lobos

Santiago - Chile
4 de noviembre de 2016

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN Y ANTECEDENTES

Propuestos en la década de los 80, los árboles de decisión son uno de los métodos de aprendizaje inductivo más usado pues son una herramienta que ayuda a la toma de decisiones de manera rápida y efectiva. En estos se evalúan las distintas alternativas intentando determinar la mejor decisión y su nombre se debe a que gráficamente el modelo adopta la forma de un árbol.

En este documento se continúa el análisis del data-set *tic-tac-toe* aplicando esta técnica esperando encontrar un patrón de juego que culmine victorioso para un jugador u otro tipo de conocimiento relevante que pueda ser aplicado por los jugadores de este pasatiempo.

1.2 OBJETIVOS

Para este laboratorio se tiene como principal objetivo extraer conocimiento del data-set tic-tac-toe utilizando árboles de decisión, para luego comparar los resultados obtenidos y los análisis pertinentes con las reglas de asociación encontradas en la experiencia anterior.

1.3 ORGANIZACIÓN DEL DOCUMENTO

El presente documento distribuye su contenido de la siguiente forma, en primer lugar se encuentra un capítulo dedicado a un pequeño marco teórico en el cual se incluye la definición de la técnica de Árbol de Decisión, de esta manera contextualizar el proceso asociado a este laboratorio.

A continuación, se presenta el capítulo de Obtención del Árbol en donde se engloba todo lo referente a los resultados dado el árbol de decisión y su proceso de obtención. En lo que sigue, se exponen los análisis de los resultados obtenidos en el desarrollo de los capítulos anteriores, y por último, con lo elaborado, se realiza una síntesis total, la cual es presentada en las conclusiones del presente documento.

CAPÍTULO 2. MARCO TEÓRICO

2.1 ÁRBOL DE DECISIÓN

Propuesta por Quinlan en el año 1983, un árbol de decisión es una técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas. Se pueden utilizar árboles de decisión para generar sistemas expertos, búsquedas binarias y árboles de juegos (Hernández Perales, s.f.). Las ventajas de un árbol de decisiones son:

- Resume los ejemplos de partida, permitiendo la clasificación de nuevos casos siempre y cuando no existan modificaciones sustanciales en las condiciones bajo las cuales se generaron los ejemplos que sirvieron para su construcción.
- Facilita la interpretación de la decisión adoptada.
- Proporciona un algo grado de comprensión del conocimiento utilizado en la toma de decisiones.
- Explica el comportamiento respecto a una determinada tarea de decisión.
- Reduce el número de variables independientes.
- Es una magnifica herramienta para el control de la gestión empresarial

Los árboles de decisión se utilizan en cualquier proceso que implique toma de decisiones (Hernández Perales, s.f.), ejemplos de estos procesos son:

- Búsqueda binaria
- Sistemas expertos
- Arboles de juego

2.2 ENTROPÍA (INFORMACIÓN)

La entropía es una medida del grado de incertidumbre asociado a una distribución de probabilidad. En una distribución uniforme, todos los valores son igualmente probables $P_i = \frac{1}{N}$ y por tanto la entropía es máxima, lo cual indica máxima incertidumbre (Cazorla Quevedo, 2011). Matemáticamente se expresa como:

$$Inf(C) = \sum_i P(C_k) ld(P(C_k)) \tag{2.1}$$

donde C es la clase.

2.3 GANANCIA DE INFORMACIÓN

La ganancia de información es una propiedad estadística que mide como clasifica el atributo analizado a los ejemplos (Malagón, 2005). En otras palabras, es una medida de cuanto ayuda el conocer el valor de

una variable aleatoria V para conocer el verdadero valor de otra C. En nuestro caso, V es un atributo de un ejemplo dado mientras que C es la clase a la que pertenece el ejemplo. Una alta ganancia implica que el atributo V permite reducir la incertidumbre de la clasificación del ejemplo de entrada (Cazorla Quevedo, 2011). Matemáticamente se expresa como:

$$Ganancia(V) = Inf(C) - Inf(C/V) \tag{2.2}$$

Con:

$$Inf(C/V) = \sum_i P(V_i) Inf(C/V_i) \tag{2.3}$$

$$Inf(C/V_i) = -\sum_i P(c_k/V_i) ld(P(c_k/v_i)) \tag{2.4}$$

2.3.1 Razón de ganancia

Existen casos en que se quieren comparar variables con distinto número de instancias, para esto de calcula la razón de ganancia la cual normaliza las ganancias para que sean comparables. Matemáticamente se expresa como:

$$RG = \frac{Ganancia(V)}{Inf(V)} \tag{2.5}$$

Con:

$$Inf(V) = -\sum_i P(V_i) ld(P(V_i)) \tag{2.6}$$

CAPÍTULO 3. OBTENCIÓN DEL ÁRBOL

Para la obtención del árbol se utilizó la función *rpart* de la biblioteca del mismo nombre. Si bien se realizaron pruebas con funciones de la biblioteca *C50* y aunque los resultados eran similares, se prefirió mostrar lo obtenido con la primera librería dado que la representación de los árboles otorga una mayor claridad de las distintas reglas a partir de una mejor visualización gráfica con las distintas funciones que se incluyen. Cabe destacar que *rpart* utiliza un algoritmo de particionamiento recursivo, al igual que el algoritmo *C5.0*, con la diferencia de los métodos de poda aplicados. Una vez aclarado el punto anterior, el procedimiento consistió en un primer paso cargar los datos, para luego crear un set de entrenamiento y uno de prueba. Con esto hecho se creó el árbol de decisión utilizando el set de entrenamiento. Una vez generadas las reglas, se aplicaron a los datos del set de prueba para realizar predicciones y así evaluar la eficiencia del árbol. Como siguiente paso se usó la función *rpart.plot* para visualizar el árbol obtenido gráficamente, este se muestra a continuación en la Figura 3.1.

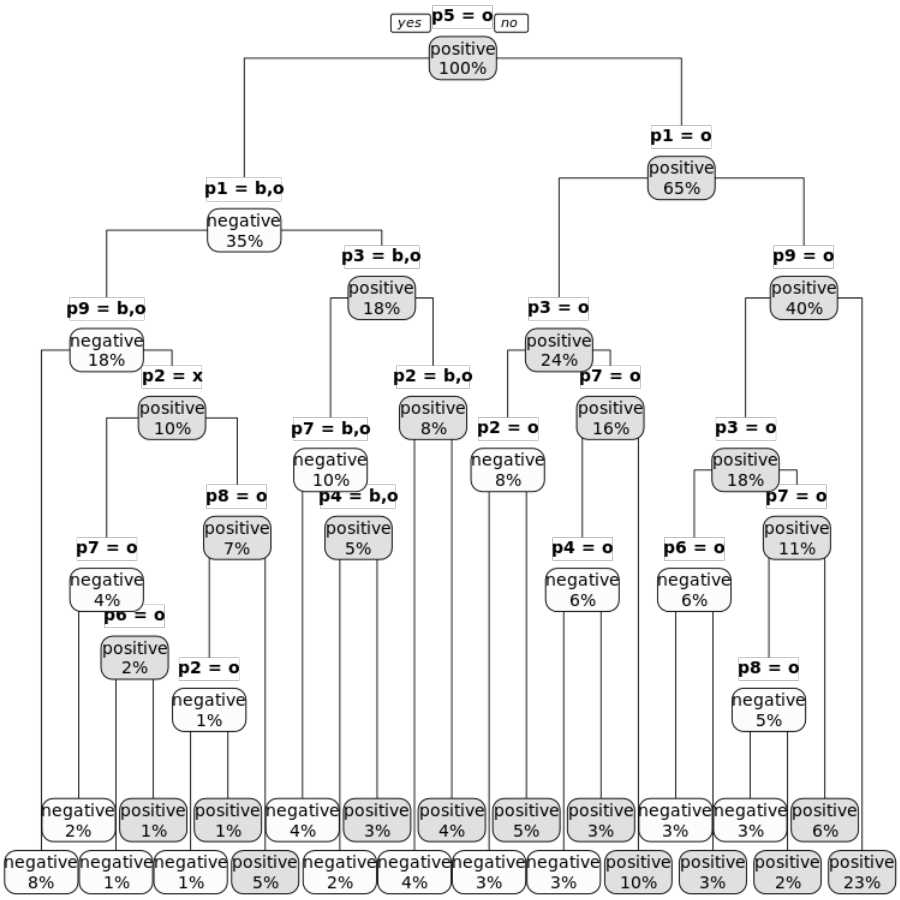

Figura 3.1: Árbol de decisión.

De manera explicativa para que se puede entender mejor la representación del árbol obtenida, las ramas que se extienden por la derecha implican el no cumplimiento de la expresión del nodo, siendo lo contrario para las ramas que se extienden por la izquierda.

Con este árbol se obtiene la siguiente tabla de predicción (Cuadro 3.1):

Cuadro 3.1: Tabla de predicción del árbol

	Negative	**Positive**
Negative	303	29
Positive	38	588

CAPÍTULO 4. ANÁLISIS DE RESULTADOS

4.1 ANÁLISIS ÁRBOL

Dado el árbol de decisión obtenido en el capítulo anterior, se pueden detallar las siguientes observaciones:

- Se puede observar que la casilla más importante en el tablero según el árbol es p5, dado que es el nodo principal, cabe recordar que esta casilla representa la posición central del tablero. Dada esta se determina en el árbol aquellas reglas que según el jugador (*x*, *o*, *b*) resultan en las dos posibles definiciones de la clase.

- Siguiendo cada rama del árbol se puede verificar que las jugadas con más probabilidad de ganar para el jugador *x* son las identificadas en experiencias anteriores las cuales incluyen en mayor proporción las diagonales y los costados. Por ejemplo la rama formada en el costado derecho del árbol muestra que con las posiciones $p5$, $p1$ y $p9$ (una de las diagonales) se obtiene un $23\,\%$ de jugadas victoriosas.

- Así mismo si se sigue la rama izquierda en donde se obtiene la misma jugada del punto anterior pero con el jugador *o*, se obtiene un porcentaje de derrotas del $8\,\%$, el más alto pues se trata de una diagonal, pero bajo en comparación al del contrincante pues es este último el que siempre inicia la partida.

- Lo mismo se puede observar en la rama $p5$ (no), $p1 = o$ (yes), $p3 = o$ (no), $p7 = o$ (no) en donde se observa que la otra diagonal tiene solo valores *x* dando un $10\,\%$ de clasificaciones positivas (el segundo porcentaje más alto con l oque las diagonales darían un $33\,\%$ de victorias al jugador *x*).

- Aquellas reglas que representan la mayor probabilidad de resultar en victoria para el jugador o, tornan en un $4\,\%$ a excepción de una de las mencionadas anteriormente que tiende al $8\,\%$. En particular una de ellas sigue la jugada diagonal ocupando las casillas $p7$, $p3$ y $p5$. En el otro caso, no es posible determinar con exactitud todas las posiciones que ocupa, pero dado que se tienen las casillas $p5$ y $p2$, si se sigue la estructura del juego y una jugada lógica, se puede estimar que se ocupa también la casilla $p8$, de manera que aquí se encuentra una posible jugada que hace ganador al jugador o que no sigue la estructura en diagonal.

- Con el árbol obtenido se tiene más de un $90\,\%$ de clasificaciones correctas según el Cuadro 3.1, dada la suma de los que Negative Negative y los Positive Positive, siendo esta la predicción dada por el árbol obtenido a través del set de entrenamiento, en base a la aplicación de este al set de prueba.

4.2 COMPARACIÓN

Para la comparación con la experiencia anterior recordaremos las reglas de asociación obtenidas.

En detalle, las reglas obtenidas con una clasificación *positive* son las siguientes:

Cuadro 4.1: Reglas obtenidas - Clase Positive

	lhs	**rhs**	**Soporte**	**Confianza**	**Lift**
[1]	p2=o	class=positive	0.2390397	0.6939394	1.061971
[2]	p6=o	class=positive	0.2390397	0.6939394	1.061971
[3]	p8=o	class=positive	0.2390397	0.6939394	1.061971
[4]	p4=o	class=positive	0.2390397	0.6939394	1.061971
[5]	p9=x	class=positive	0.3079332	0.7057416	1.080033
[6]	p7=x	class=positive	0.3079332	0.7057416	1.080033
[7]	p3=x	class=positive	0.3079332	0.7057416	1.080033
[8]	p1=x	class=positive	0.3079332	0.7057416	1.080033
[9]	p5=x	class=positive	0.3820459	0.7991266	1.222945

Para la clasificación *negative* se presentaron las reglas dadas a continuación:

Cuadro 4.2: Reglas obtenidas - Clase Negative

	lhs	**rhs**	**Soporte**	**Confianza**	**Lift**
[1]	p7=o	class=negative	0.1524008	0.4358209	1.257580
[2]	p9=o	class=negative	0.1524008	0.4358209	1.257580
[3]	p3=o	class=negative	0.1524008	0.4358209	1.257580
[4]	p1=o	class=negative	0.1524008	0.4358209	1.257580
[5]	p5=o	class=negative	0.2004175	0.5647059	1.629483
[6]	p8=x	class=negative	0.1597077	0.4047619	1.167958
[7]	p6=x	class=negative	0.1597077	0.4047619	1.167958
[8]	p4=x	class=negative	0.1597077	0.4047619	1.167958
[9]	p2=x	class=negative	0.1597077	0.4047619	1.167958

En relación a las reglas y considerando el árbol de decisión podemos observar que:

- Como ya se dijo, las casillas más utilizadas para lograr una jugada ganadora son principalmente la central y las esquinas del tablero, esto se puede observar tanto en el árbol (ver sección anterior) como en las reglas de asociación en donde los valores de *lift* mayores son los obtenidos por estas casillas cuando la clasificación es positiva (Cuadro 4.1).
- También se ratifica que con estas variables es con las que se generan más resultados ganadores siendo estas las combinaciones dichas en el análisis del árbol y en el caso de las reglas de asociación las

combinaciones de los con lift mayor incluyendo $p5$ (Cuadro 4.1).

- Aunque no se analizaron los casos de las casillas $p2$, $p4$, $p6$ y $p8$ por tener bajos porcentajes de clasificación, se puede observar que en las reglas de asociación son las con menor *lift* con lo que no nos entregan mayor información, pero conociendo el contexto del data-set y la jugabilidad del *tic tac toe* se puede inferir que la razón de estos bajos valores es que se trata de las casillas con menos posibles jugadas ganadoras (solo 2 cada una). Esto se verifica en cada experiencia realizada.

- Un caso particular que no se refleja en las reglas referenciadas y que se pudo dilucidar a través del árbol, es que existe una jugada que es vertical y ocupa las casillas centrales del tablero, es decir, $p2$, $p5$ y $p8$, la cual resulta con ganador al jugador o, siendo una de las reglas con mayor probabilidad (aunque muy baja, 4 %) dentro del ámbito de la clase negative.

CAPÍTULO 5. CONCLUSIONES

Luego de analizar el árbol de decisión podemos destacar que con este método, gracias a su forma, es mucho más simple visualizar las instancias de las variables que aparecen en los respectivos nodos. Esto es muy importante desde el punto de vista del investigador pues, en el caso del *tic tac toe* se puede construir una jugada que calce con lo arrojado por el método y así evaluar si la jugada está o no bien clasificada (aunque como se vio en el Cuadro 3.1, el modelo tiene un pequeño porcentaje de error a considerar al momento de verificar una jugada). También se debe considerar que la jugada puede construirse, pero el camino que sigue una rama del árbol no necesariamente dice el orden en el que se llena el tablero pues de ser así, se estarían obviando muchos casos; aún así, en general se verificaron manualmente distintos caminos y la clasificación final correspondía en su mayoría a la mostrada en la Figura 3.1.

En cuanto a la comparación realizada con las reglas de asociación obtenidas en la experiencia anterior se pudo observar una clara tendencia a que la casilla central junto a las cuatro esquinas culminan en jugadas ganadoras tal y como se venía viendo no solo en estas 2 experiencias sino que desde antes. Además se hizo hincapié en las casillas restantes ($p2$, $p4$, $p6$ y $p8$), de las que se determinó que los bajos porcentajes de clasificación y bajos valores para *lift* en el caso de las reglas de asociación, se debe a que al tener menos posibles combinaciones ganadoras (cada una con solo 2), la cantidad de jugadas con clasificación positive son menos y en su mayoría son clasificadas con una derrota para *x*; esta hipótesis que se venía dando desde la primera experiencia está más que ratificada con los distintos métodos aplicados al data-set.
A pesar de la rectificación del conocimiento obtenido con ambas técnicas, se reitera la idea de que hubiera sido mejor en la búsqueda de dicho conocimiento, tener los órdenes de las jugadas realizadas en cada instancia de manera que se pudiera encontrar patrones de juego en base a esos órdenes.

Cabe recordar que las reglas de asociación permiten expresar patrones de comportamiento entre los datos en función de su aparición conjunta, expresando las combinaciones de valores de los atributos que ocurren mas veces, así mismo, un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, en donde la decisión final depende del camino que se toma desde la raíz del árbol. *(Pérez, s.f.)*. Entonces, la principal diferencia que se denota es que los arboles usan la heurística de evaluación sobre un atributo y normalmente realizan un sobreajuste seguido del podado, en cambio las reglas de asociación consideran cualquier conjunto de atributos con cualquier otro conjunto de atributos, basándose en en medidas de confianza y soporte. A partir de ello y con lo observado durante el desarrollo de esta experiencia, a pesar de que los análisis pertinentes infieren un conocimiento similar para ambas técnicas, el proceso de obtención de dicho conocimiento es diferente dado los resultado obtenidos. Es así como para el análisis de las reglas de asociación se tomaron los distintos atributos que presentaban un mayor soporte y confianza determinando la clase en particular, en cambio, para el análisis del árbol de decisión, se partió de la raíz para la evaluación de cada rama, llegando finalmente a la definición de cada clase según el recorrido hecho.

CAPÍTULO 6. REFERENCIAS

Cazorla Quevedo, M. (2011). Aprendizaje: arboles de decisión. Recuperado desde https://rua.ua.es/dspace/bitstream/10045/17323/10/Aprendizaje.arboles.pdf

Hernández Perales, J. (s.f.). Árbol de decisión. Recuperado desde http://www.utm.mx/~jahdezp/archivos%20estructuras/DESICION.pdf

Malagón, C. (2005). Aprendizaje automático mediante árboles de decisión. Recuperado desde https://www.nebrija.es/~cmalagon/inco/apuntes_mios/arboles_de_decision.pdf

Package 'c50'. (2015). Recuperado desde https://cran.r-project.org/web/packages/C50/C50.pdf

Pérez, S. L. P. (s.f.). Minería de datos (reglas de asociación y árboles de decisión). Recuperado desde http://computacion.cs.cinvestav.mx/~sperez/cursos/md/14i/ReglasAsociacionYArboles.pdf

Terry Therneau, B. R., Beth Atkinson. (2015). Package 'rpart'. Recuperado desde https://cran.r-project.org/web/packages/rpart/rpart.pdf

CAPÍTULO 7. ANEXO: CÓDIGO FUENTE EN R

```
library(C50)
library(rpart)
library(rpart.plot)
DB <- read.table("Lab5/tic-tac-toe.data", header=FALSE, sep=",",
      col.names=c("p1","p2","p3","p4","p5","p6","p7","p8","p9","class"))

#arbol1<- C50::C5.0(class ~., data = DB)
#summary(arbol1)

#arbol2<- C50::C5.0(class ~., data = DB,rules = TRUE)
#summary(arbol2)



#http://apuntes-r.blogspot.cl/2014/09/predecir-perdida-de-clientes-con-arbol.html
training    <- DB
test        <- DB

Arbol1<-rpart(class ~ .,data=training,parms=list(split="information"),
        method = "class", control = rpart.control(cp = 0.01))

Prediccion <- predict(Arbol1, test,type="class")
MC         <- table(test[, "class"], Prediccion)
print(MC)

par(mar = rep(2, 4))
rpart.plot(Arbol1, type=1, extra=100,cex = .7,
           box.col=c("gray99", "gray88")[Arbol1$frame$yval])

#summary(Arbol1)
```