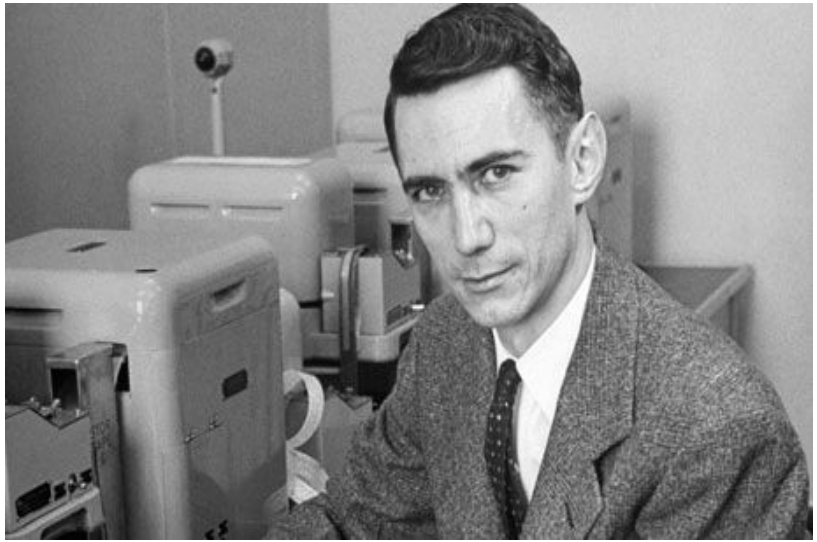
Teoría de la Información
Árboles de decisión

Trabajo de Historia de la Computación

María Oliver Balsalobre, Ramón Casado Gómez, Miguel Sánchez Maldonado

Índice

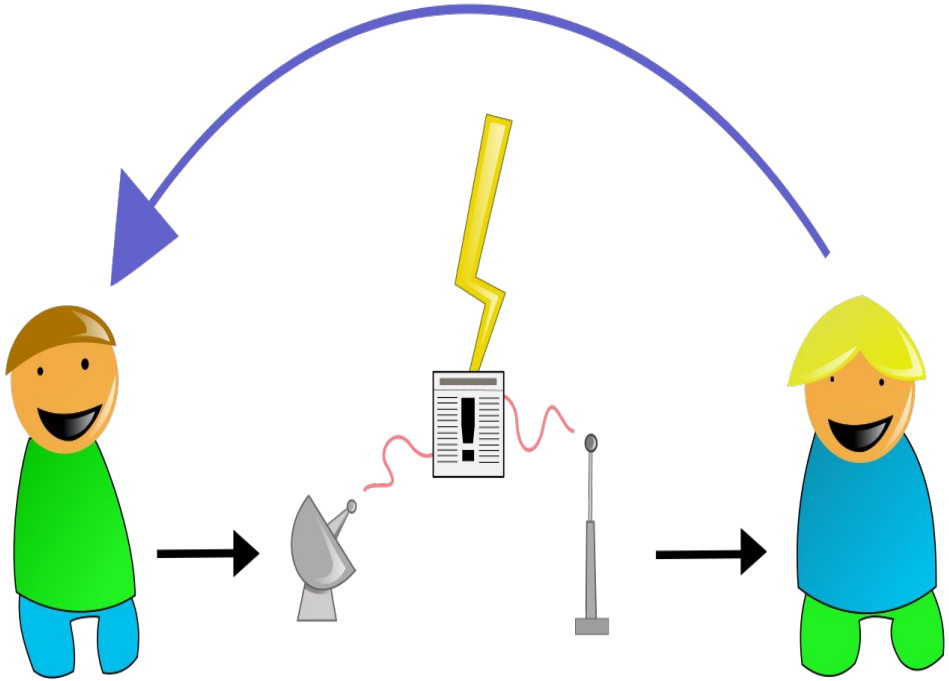
1. Introducción



La teoría de la información surgió a finales de la Segunda Guerra Mundial, gracias, sobre todo, a Claude Elwood Shannon, conocido como "el padre de la teoría de la información".

Se tenía la necesidad de reducir la cantidad de datos sin perder información, por lo que muestra el camino a seguir para determinar la cantidad de información de manera que los datos sean representados de una manera eficiente.

1. Introducción



Está relacionada con las leyes matemáticas que rigen la transmisión y el procesamiento de la información y se ocupa de la medición de la información y de su representación. Así pues, es una rama de la teoría matemática y de las ciencias de la computación.

Índice

2. Entropía

Nos permite determinar cuánta información aporta una fuente de datos.
La teoría de Shannon nos dice: más información = menos entropía

La entropía es una cota inferior para cualquier codificación.

2. Entropía

El cálculo de la entropía viene dado por la siguiente fórmula:

$$\sum_{v \in Valores(A)} \frac{|D_v|}{|D|} Ent(D_v)$$

D es el conjunto total de ejemplos y Dv es el subconjunto de ejemplos de D con valor del atributo A igual a v.
Cuanta más homogénea sea la distribución de probabilidades, menor será la entropía y no podemos encontrar ningún método que codifique información con menos bits de los que marca la entropía,

2. Entropía

Veamos la fórmula de la entropía para un caso particular para dos posibles sucesos.

$$Ent\,(D) = (-\frac{|P|}{|D|}log_2\frac{|P|}{|D|} - \frac{|N|}{|D|}log_2\frac{|N|}{|D|})$$

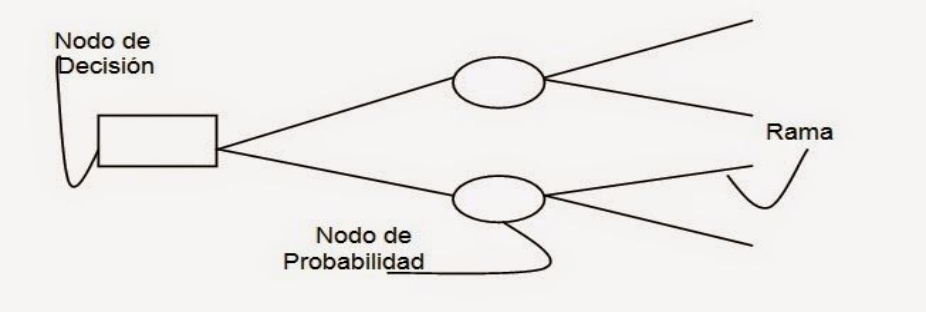
Donde P serán los subconjuntos de ejemplos positivos de D y N los subconjuntos de ejemplos negativos de D.

Índice

3. ÁRBOLES DE DECISIÓN

Son modelos de predicción utilizados en el ámbito de la IA. Son una forma gráfica y visual de representar todos los sucesos que pueden surgir a partir de una decisión, por lo que nos ayudarán a elegir la opción más acertada desde un punto de vista probabilístico.

Lleva a cabo un test a medida que se recorre hacia las hojas del árbol para alcanzar una decisión.



3. Ejemplo

Veamos un ejemplo en el que nos preguntamos si Salir de fiesta o no.

3. Ejemplo.

Veamos el mismo ejemplo de Salir de fiesta o No cambiando el orden de las condiciones

Índice

4. Algoritmo ID3 (Induction Decision Trees)

- Fue desarrollado por J.Ross Quinlan en 1983.
- Objetivo: construir un árbol de decisión que explique cada instancia (celda) de la secuencia de entrada de la manera más compacta posible a partir de una tabla de inducción.
- Elige el mejor atributo dependiendo de una determinada heurística.
- Determina las variables que contienen información relevante para la solución del sistema.

4. Algoritmo ID3 (Induction Decision Trees)

Características:

- Recursividad: seguir aplicando algoritmo al siguiente subárbol.
- Escoge el mejor atributo
- Es capaz de llegar a nodos hoja
- No hace uso de backtracking
- Utiliza la entropía

4. EJEMPLO

Tabla de inducción

	Cielo	Temperatura	Humedad	Viento	Posibilidad de jugar al tenis
Día 1	Sol	Alta	Alta	Débil	-
Día 2	Sol	Alta	Alta	Fuerte	-
Día 3	Nubes	Alta	Alta	Débil	+
Día 4	Lluvia	Suave	Alta	Débil	+
Día 5	Lluvia	Baja	Normal	Débil	+
Día 6	Lluvia	Baja	Normal	Fuerte	-
Día 7	Nubes	Baja	Normal	Fuerte	+
Día 8	Sol	Suave	Alta	Débil	-
Día 9	Sol	Baja	Normal	Débil	+
Día 10	Lluvia	Suave	Normal	Débil	+
Día 12	Sol	Suave	Normal	Fuerte	+
Día 13	Nubes	Suave	Alta	Fuerte	+
Día 14	Nubes	Alta	Normal	Débil	+
Día 15	Lluvia	Suave	Alta	Fuerte	-

	Cielo	Temperatura	Humedad	Viento	Posibilidad de jugar al tenis
Día 1	Sol	Alta	Alta	Débil	-
Día 2	Sol	Alta	Alta	Fuerte	-
Día 3	Nubes	Alta	Alta	Débil	+
Día 4	Lluvia	Suave	Alta	Débil	+
Día 5	Lluvia	Baja	Normal	Débil	+
Día 6	Lluvia	Baja	Normal	Fuerte	-
Día 7	Nubes	Baja	Normal	Fuerte	+
Día 8	Sol	Suave	Alta	Débil	-
Día 9	Sol	Baja	Normal	Débil	+
Día 10	Lluvia	Suave	Normal	Débil	+
Día 12	Sol	Suave	Normal	Fuerte	+
Día 13	Nubes	Suave	Alta	Fuerte	+
Día 14	Nubes	Alta	Normal	Débil	+
Día 15	Lluvia	Suave	Alta	Fuerte	-

4. EJEMPLO

Aplicamos el algoritmo ID3 y utilizamos la fórmula de la entropía:

$$\sum_{v \in Valores(A)} \frac{|D_v|}{|D|} Ent(D_v)$$

Además veremos cuál será la entropía esperada después de elegir el mejor atributo, calculándola para el resto de ejemplos.

4. EJEMPLO

- Sol: $Ent([2+,3-]) = -\frac{2}{14}\cdot log_2(\frac{2}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 0'97$
- Nubes: $Ent([4+,0-]) = 0$
- Lluvia: $Ent([2+,3-]) = -\frac{2}{14}\cdot log_2(\frac{2}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 0'97$

$$Entropia(D,Cielo) = \frac{5}{14}\cdot 0'97 + \frac{5}{14}\cdot 0'97 = 0'7$$

	Cielo	Posibilidad de jugar al tenis
Día 1	Sol	-
Día 2	Sol	-
Día 3	Nubes	+
Día 4	Lluvia	+
Día 5	Lluvia	+
Día 6	Lluvia	-
Día 7	Nubes	+
Día 8	Sol	-
Día 9	Sol	+
Día 10	Lluvia	+
Día 12	Sol	+
Día 13	Nubes	+
Día 14	Nubes	+
Día 15	Lluvia	-

4. EJEMPLO

- Alta: $Ent([2+,2-]) = -\frac{2}{14}\cdot log_2(\frac{2}{14}) - \frac{2}{14}\cdot log_2(\frac{2}{14}) = 1$
- Suave: $Ent([4+,2-]) = -\frac{4}{14}\cdot log_2(\frac{4}{14}) - \frac{2}{14}\cdot log_2(\frac{2}{14}) = 0'91$
- Baja: $Ent([3+,1-]) = -\frac{3}{14}\cdot log_2(\frac{3}{14}) - \frac{1}{14}\cdot log_2(\frac{1}{14}) = 0'81$

$$Entropía(D, Temperatura) = \frac{4}{14}\cdot 1 + \frac{6}{14}\cdot 0'91 + \frac{4}{14}\cdot 0'81 = 0'92$$

	Temperatura	Posibilidad de jugar al tenis
Día 1	Alta	-
Día 2	Alta	-
Día 3	Alta	+
Día 4	Suave	+
Día 5	Baja	+
Día 6	Baja	-
Día 7	Baja	+
Día 8	Suave	-
Día 9	Baja	+
Día 10	Suave	+
Día 12	Suave	+
Día 13	Suave	+
Día 14	Alta	+
Día 15	Suave	-

4. EJEMPLO

- Alta: $Ent([3+,4-]) = -\frac{3}{14}\cdot log_2(\frac{3}{14}) - \frac{4}{14}\cdot log_2(\frac{4}{14}) = 0'99$
- Normal: $Ent([6+,1-]) = -\frac{6}{14}\cdot log_2(\frac{6}{14}) - \frac{1}{14}\cdot log_2(\frac{1}{14}) = 0'59$

$$Entropía(D, Humedad) = \frac{7}{14}\cdot 0'99 + \frac{7}{14}\cdot 0'59 = 0'79$$

	Humedad	Posibilidad de jugar al tenis
Día 1	Alta	-
Día 2	Alta	-
Día 3	Alta	+
Día 4	Alta	+
Día 5	Normal	+
Día 6	Normal	-
Día 7	Normal	+
Día 8	Alta	-
Día 9	Normal	+
Día 10	Normal	+
Día 12	Normal	+
Día 13	Alta	+
Día 14	Normal	+
Día 15	Alta	-

4. EJEMPLO

- Fuerte: $Ent([6+,2-]) = -\frac{6}{14}\cdot log_2(\frac{6}{14}) - \frac{2}{14}\cdot log_2(\frac{2}{14}) = 0'81$
- Débil: $Ent([3+,3-]) = -\frac{3}{14}\cdot log_2(\frac{3}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 1$

$$Entropía(D, Viento) = \frac{8}{14}\cdot 0'81 + \frac{6}{14}\cdot 1 = 0'9$$

	Viento	Posibilidad de jugar al tenis
Día 1	Débil	-
Día 2	Fuerte	-
Día 3	Débil	+
Día 4	Débil	+
Día 5	Débil	+
Día 6	Fuerte	-
Día 7	Fuerte	+
Día 8	Débil	-
Día 9	Débil	+
Día 10	Débil	+
Día 12	Fuerte	+
Día 13	Fuerte	+
Día 14	Débil	+
Día 15	Fuerte	-

4. EJEMPLO

$$Entropía(D, Cielo) = \frac{5}{14} \cdot 0'97 + \frac{5}{14} \cdot 0'97 = 0'7$$

$$Entropía(D, Temperatura) = \frac{4}{14} \cdot 1 + \frac{6}{14} \cdot 0'91 + \frac{4}{14} \cdot 0'81 = 0'92$$

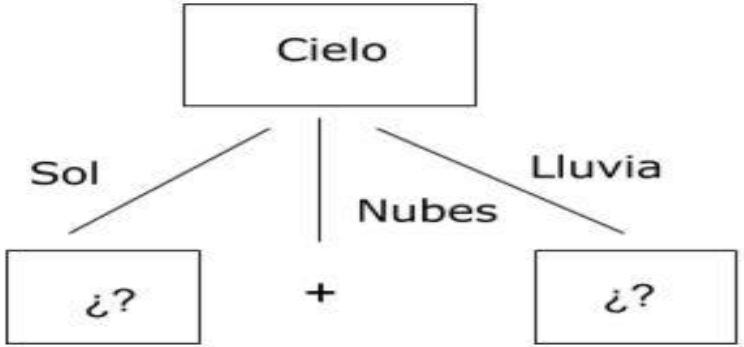
$$Entropía(D, Humedad) = \frac{7}{14} \cdot 0'99 + \frac{7}{14} \cdot 0'59 = 0'79$$

$$Entropía(D, Viento) = \frac{8}{14} \cdot 0'81 + \frac{6}{14} \cdot 1 = 0'9$$

Como vemos, el que produce menos entropía es Cielo.
Por tanto, éste será el mejor atributo,

4. EJEMPLO

Por tanto, como Cielo es el mejor atributo, el árbol de decisión empezaría de la siguiente manera.



En el caso de que el día esté nublado, hay 4 subconjuntos de ejemplos positivos y ninguno negativo, por lo que claramente sí será posible jugar al tenis.

4. EJEMPLO

	Cielo	Temperatura	Humedad	Viento	Posibilidad de jugar al tenis
Día 1	Sol	Alta	Alta	Débil	-
Día 2	Sol	Alta	Alta	Fuerte	-
Día 8	Sol	Suave	Alta	Débil	-
Día 9	Sol	Baja	Normal	Débil	+
Día 12	Sol	Suave	Normal	Fuerte	+

- $Entropía(D_{sol}, Temperatura) = \frac{2}{5}\cdot 0 + \frac{2}{5}\cdot 1 + \frac{1}{5}\cdot 0 = 0'4$
- $Entropía(D_{sol}, Humedad) = \frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0 = 0$
- $Entropía(D_{sol}, Viento) = \frac{2}{5}\cdot 1 + \frac{3}{5}\cdot 0'91 = 0'96$

4. EJEMPLO

	Cielo	Temperatura	Humedad	Viento	Posibilidad de jugar al tenis
Día 4	Lluvia	Suave	Alta	Débil	+
Día 5	Lluvia	Baja	Normal	Débil	+
Día 6	Lluvia	Baja	Normal	Fuerte	-
Día 10	Lluvia	Suave	Normal	Débil	+
Día 15	Lluvia	Suave	Alta	Fuerte	-

- $Entropía(D_{Lluvia}, Temperatura) = \frac{3}{5} \cdot 0'91 + \frac{2}{5} \cdot 1 = 0'15$
- $Entropía(D_{Lluvial}, Humedad) = \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot 0'91 = 0'15$
- $Entropía(D_{Lluvia}, Viento) = \frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0 = 0$

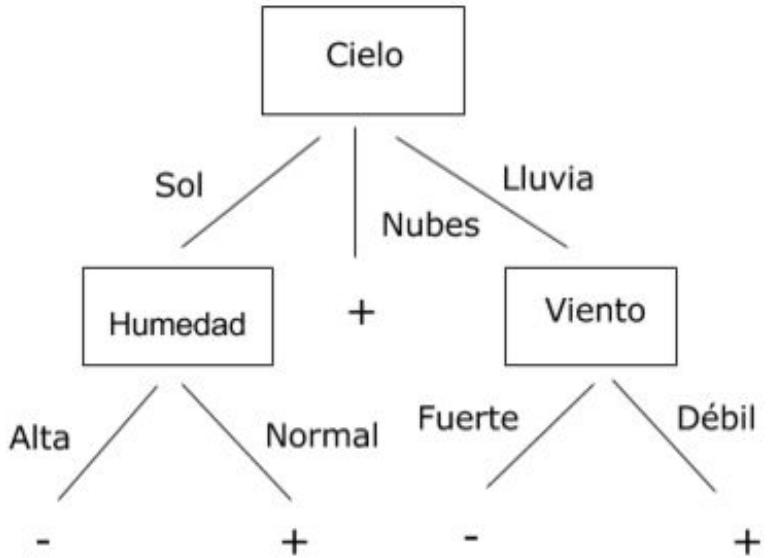
4. EJEMPLO

	Cielo	Humedad	Posibilidad de jugar al tenis
Día 1	Sol	Alta	-
Día 2	Sol	Alta	-
Día 8	Sol	Alta	-
Día 9	Sol	Normal	+
Día 12	Sol	Normal	+

	Cielo	Viento	Posibilidad de jugar al tenis
Día 4	Lluvia	Débil	+
Día 5	Lluvia	Débil	+
Día 6	Lluvia	Fuerte	-
Día 10	Lluvia	Débil	+
Día 15	Lluvia	Fuerte	-

4. EJEMPLO

Ya tenemos información suficiente para construir el árbol de decisión minimal:



¡No se tiene en cuenta la condición Temperatura!

4. Inconvenientes algoritmo ID3

A la vista del ejemplo, podemos darnos cuenta de que el algoritmo ID3 presenta algunos inconvenientes:

- Favorece a los atributos con muchos valores, que no tienen porqué ser los más útiles.
- Datos Perdidos
- Genera grandes árboles que no representan garantía de reglas eficientes.

4. Algoritmos C4.5 y C5.0

Para intentar solucionar estos inconvenientes del algoritmo ID3, surgen más adelante otros dos algoritmos C4.5 y C5.0, también desarrollados por Quinlan, que realizan una serie de cambios tales como:

- Manejo de datos perdidos.
- Posibilidad de trabajar con datos continuos.
- Poder usar pre-poda y post-poda.
- Mejorar la eficiencia computacional.
- Árboles menos frondosos.

Índice

5. Aplicaciones de la teoría de la información

Cualquiera en las que se utilicen árboles de decisión (más extensos que los del ejemplo) los cuales permiten, de forma visual, analizar de manera fácil los casos que queramos estudiar.

- Compresión con/sin pérdida de datos.
- Codificación de canal.
- Criptografía.
- Comprensión de agujeros negros.
- Factibilidad de teléfonos móviles.
- Sistemas expertos como Mycin.

FIN DE LA PRESENTACIÓN