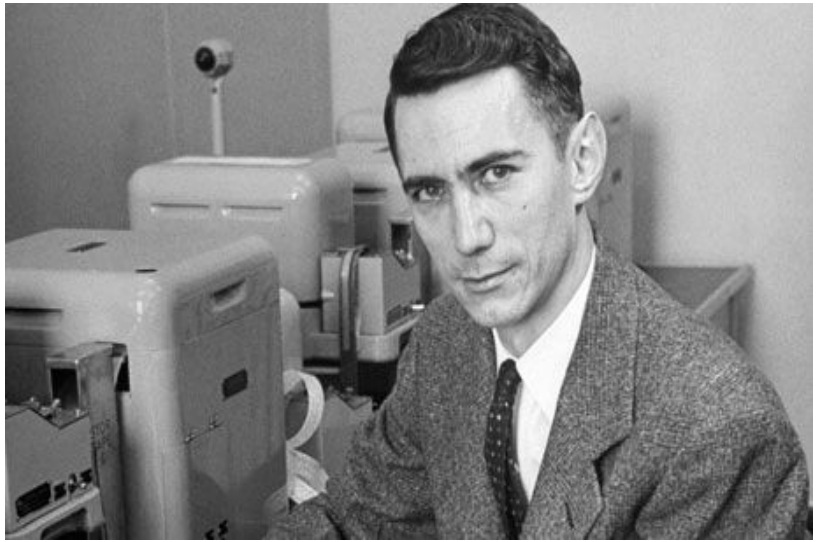
Teoría de la Información
Árboles de decisión

Trabajo de Historia de la Computación

María Oliver Balsalobre, Ramón Casado Gómez, Miguel Sánchez Maldonado

Índice

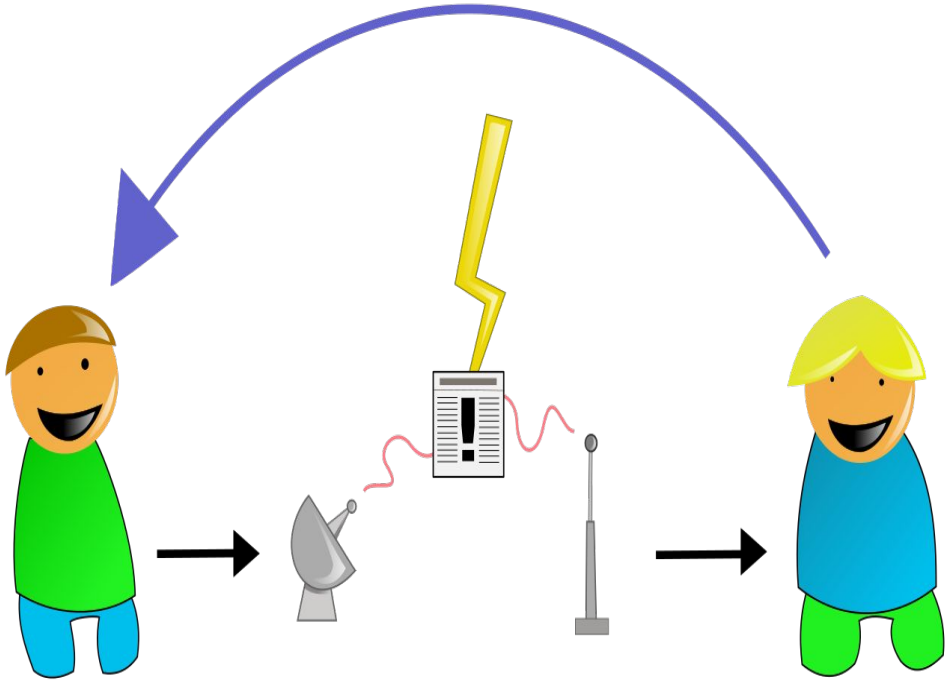
1. Introducción



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

Mide la información aportada por una fuente de datos. Es decir, ante una determinada situación, es la probabilidad de que ocurra cada uno de los resultados posibles.

La teoría de Shannon nos dice que la información reduce la incertidumbre o entropía.

2. Entropía

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

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

Cuanta más homogénea sea la distribución de probabilidades, mayor será la entropía y no podemos encontrar ningún método que codifique información con menos bits de los que marca la entropía,

Índice

3. Algoritmo ID3 (Induction Decision Trees)

- Fue desarrollado por J.Ross Quinlan en 1983.
- Objetivo: construir un árbol de decisión que explique cada instancia 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.

3. Algoritmo ID3 (Induction Decision Trees)

Características:

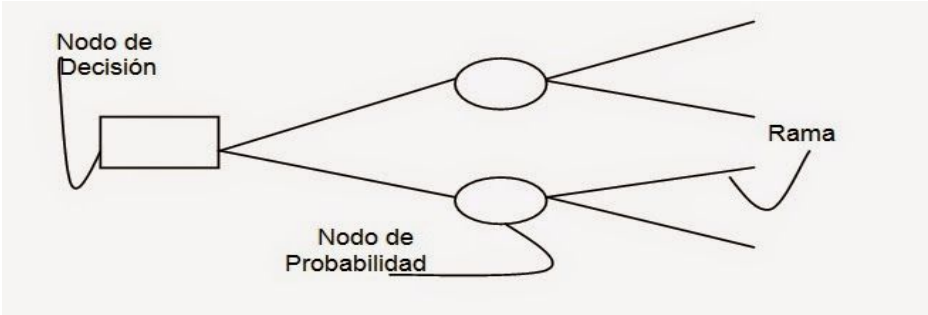
- Recursividad
- Escoge el mejor atributo
- Es capaz de llegar a nodos hoja
- No hace uso de backtracking
- Utiliza la entropía
- Usa el concepto de ganancia de información para seleccionar el atributo más útil en cada paso.

Índice

4. 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.



4. EJEMPLO

	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:

- D es el conjunto de ejemplos (días).
- P son los subconjuntos de ejemplos positivos.
- N son los subconjuntos de ejemplos negativos.

Además calcularemos cuáles serán la entropía esperada y la ganancia de información, después de usar un atributo A en el árbol:

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

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

4. EJEMPLO

- **Cielo**: $Ent([9+,5-]) = -\frac{9}{14}\cdot log_2(\frac{9}{14}) - \frac{5}{14}\cdot log_2(\frac{5}{14}) = 0'94$
  - 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: $Ent([2+,3-]) = -\frac{2}{14}\cdot log_2(\frac{2}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 0'97$

$Ganancia(D, Cielo) = 0'94 - \frac{5}{14}\cdot 0'97 - \frac{5}{14}\cdot 0'97 = 0'24$

- **Temperatura**: $Ent([9+,5-]) = -\frac{9}{14}\cdot log_2(\frac{9}{14}) - \frac{5}{14}\cdot log_2(\frac{5}{14}) = 0'94$
  - 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: $Ent([3+,1-]) = -\frac{3}{14}\cdot log_2(\frac{3}{14}) - \frac{1}{14}\cdot log_2(\frac{1}{14}) = 0'81$

$Ganancia(D, Temperatura) = 0'94 - \frac{4}{14}\cdot 1 - \frac{6}{14}\cdot 0'91 - \frac{4}{14}\cdot 0'81 = 0'02$

- **Humedad**: $Ent([9+,5-]) = -\frac{9}{14}\cdot log_2(\frac{9}{14}) - \frac{5}{14}\cdot log_2(\frac{5}{14}) = 0'94$
  - 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: $Ent([6+,1-]) = -\frac{6}{14}\cdot log_2(\frac{6}{14}) - \frac{1}{14}\cdot log_2(\frac{1}{14}) = 0'59$

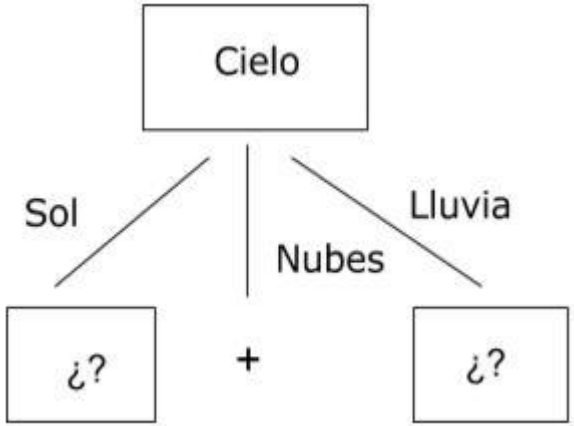
$Ganancia(D, Humedad) = 0'94 - \frac{7}{14}\cdot 0'99 - \frac{7}{14}\cdot 0'59 = 0'15$

- **Viento**: $Ent([9+,5-]) = -\frac{9}{14}\cdot log_2(\frac{9}{14}) - \frac{5}{14}\cdot log_2(\frac{5}{14}) = 0'94$
  - 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: $Ent([3+,3-]) = -\frac{3}{14}\cdot log_2(\frac{3}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 1$

$Ganancia(D, Viento) = 0'94 - \frac{8}{14}\cdot 0'81 - \frac{6}{14}\cdot 1 = 0'04$

4. EJEMPLO

Como vemos, el que produce mayor ganancia de información, es Cielo.
Por tanto, éste será el mejor atributo, luego el árbol de decisión empezaría de la siguiente manera:



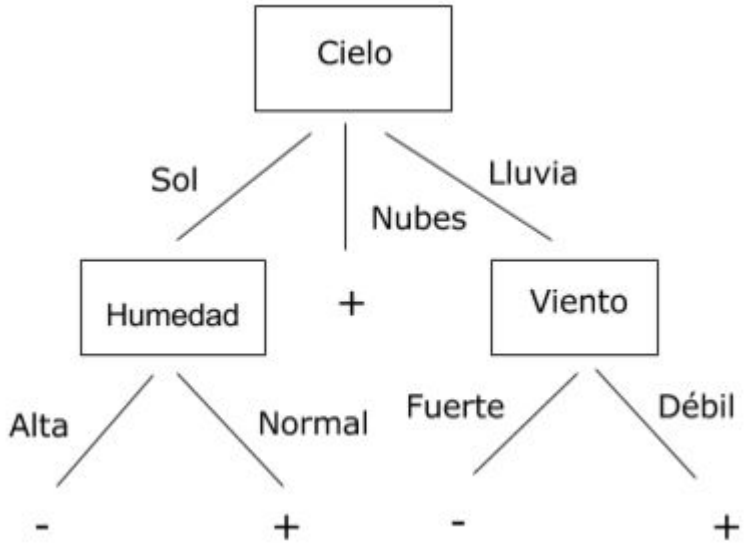
En el caso de que el día esté nublado, hay 4 subconjuntos de ejemplos positivos y ninguno negativo, por lo que claramente sí será posible jugar al tenis.

4. EJEMPLO

Estudiamos los otros dos casos para el nodo Cielo:

- Sol: $Ent([2+,3-]) = -\frac{2}{14}\cdot log_2(\frac{2}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 0'97$
  - $Ganancia(D_{sol}, Temperatura) = 0'97 - \frac{2}{5}\cdot 0 - \frac{2}{5}\cdot 1 - \frac{1}{5}\cdot 0 = 0'57$
  - $Ganancia(D_{sol}, Humedad) = 0'97 - \frac{3}{5}\cdot 0 - \frac{2}{5}\cdot 0 = 0'97$ (mejor)
  - $Ganancia(D_{sol}, Viento) = 0'97 - \frac{2}{5}\cdot 1 - \frac{3}{5}\cdot 0'91 = 0'01$
- Lluvia: $Ent([2+,3-]) = -\frac{2}{14}\cdot log_2(\frac{2}{14}) - \frac{3}{14}\cdot log_2(\frac{3}{14}) = 0'97$
  - $Ganancia(D_{sol}, Temperatura) = 0'97 - \frac{3}{5}\cdot 0'91 - \frac{2}{5}\cdot 1 = 0'82$
  - $Ganancia(D_{sol}, Humedad) = 0'97 - \frac{2}{5}\cdot 1 - \frac{3}{5}\cdot 0'91 = 0'82$
  - $Ganancia(D_{sol}, Viento) = 0'97 - \frac{3}{5}\cdot 0 - \frac{2}{5}\cdot 0 = 0'97$ (mejor)

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



4. EJEMPLO

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.
- Diferentes soluciones se alcanzan con variables con los mismos valores asociados.
- Genera grandes árboles que no representan garantía de reglas eficientes.

4. EJEMPLO

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

Cualquiera en las que se utilicen árboles de decisión:

- Compresión con/sin pérdida de datos.
- Codificación de canal.
- Criptografía.
- Estudio de la lingüística y percepción humana.
- Comprensión de agujeros negros.
- Factibilidad de teléfonos móviles.

Además, los árboles de decisión (más extensos que los del ejemplo) permiten, de forma visual, analizar de manera fácil los casos que queramos estudiar.