CAPÍTULO 3

MARCO TEÓRICO

En este capítulo se hace una descripción de la teoría para realizar una heurística de clasificación. En la sección 3.1 se describe la teoría de la información, la cual trata al término "información" como una cantidad mensurable, se habla del concepto de entropía de Shannon y cómo se define matemáticamente y en la sección 3.2 se introduce el término de ganancia de la información y la heurística para clasificar información.

3.1 TEORIA DE LA INFORMACIÓN

Esta teoría fue inicializada por Claude Shannon en 1948, es una rama de la probabilidad, en la cual se propone e investiga un nuevo modelo matemático de sistemas de comunicación. Una de las aportaciones más importantes de este modelo es tratar a los componentes de un sistema de comunicación (canales de comunicación, códigos, etc.) como entidades de probabilidad. La investigación de Shannon se basaba en el problema de la transmisión eficiente de la información [15].

Uno de los postulados básicos de la teoría de la información es que la "información" se puede tratar como una cantidad física mensurable, tal como densidad o masa. La teoría se ha aplicado extensamente por los ingenieros de comunicación y algunos de sus conceptos son utilizados en la psicología y la lingüística [15].

Dentro de esta teoría Shannon definió que cuando se calcula un valor numérico para la "información", ésta siempre es una capacidad y sugirió que esta capacidad puede

cuantificarse introduciendo así, el concepto de entropía. Entropía según la segunda ley de la termodinámica, es el grado de desorden o aleatoriedad en un sistema. Aplicando el concepto para esta teoría la Entropía de Shannon es una medida de la información o incertidumbre de experimento probabilísticos [5].

En la teoría de probabilidad, un sistema completo de eventos $A_1, A_2, A_3, \ldots, A_n$ significa un conjunto de eventos tal que uno y sólo uno de estos puede ocurrir en cada ensayo. Sí se dan los eventos $A_1, A_2, A_3, \ldots, A_n$ de un sistema completo, junto con sus probabilidades $p_1, p_2, p_3, \ldots p_n \left(p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right)$, entonces se dice que se tiene un esquema finito A [8].

$$A = \begin{pmatrix} A_1 A_2 \ldots A_n \\ p_1 p_2 \cdots p_n \end{pmatrix} \tag{3.1}$$

Cada esquema finito describe un estado de incertidumbre. Se tiene un experimento, el resultado de éste debe ser uno de los eventos $A_1, A_2, A_3, \ldots, A_n$ y se conoce solo la probabilidad de que ocurra este evento. La cantidad de incertidumbre es diferente en diferentes esquemas [8]. Esto se demuestra en estas dos alternativas

$$\begin{pmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{pmatrix} \qquad \begin{pmatrix} A_1 & A_2 \\ 0.99 & 0.01 \end{pmatrix} \tag{3.2}$$

La primera alternativa obviamente representa mucho más incertidumbre que la segunda; en el segundo caso, el resultado del experimento de A_1 es más certero, mientras que en el primer caso naturalmente se debe abstener de hacer cualquier predicción [8].

Para muchas aplicaciones parece deseable introducir una cantidad la cual proporcione una manera razonable de medir la cantidad de incertidumbre asociada con un esquema finito dado.

$$H(p_1, p_2, ..., p_n) = -K\sum_{i=1}^{n} p_i \log p_i \tag{3.3}$$

La cantidad puede servir como una medida deseable de la incertidumbre del esquema finito (3.1). La ecuación 3.3 representa la definición de la Entropía de Shannon. La base logarítmica puede ser arbitraria, la justificación para esta arbitrariedad es que si se toma el logaritmo en base 2, es porque la incertidumbre en el esquema consiste en dos eventos con igual probabilidad de ocurrir (Sí ó No) y la unidad de entropía es llamada Bit. Cuando se utilizan logaritmos de base 10, son llamados Hartley. Bits son las unidades generalmente utilizadas [14].

Para introducir el término de entropía de Shannon en un sistema de clasificación, se realizó un ejemplo para clasificar a las clases C_1 a C_5 que contienen los elementos del 1 al 20 y son descritas por 5 atributos (A_1 a A_5) los cuales pueden tomar valores de 1 o 0 (Ver Tabla 3.1).

Tabla 3.1 Agrupación de los elementos (1, 2, 3,...) en clases (C_1, C_2, C_3,…) y por tipo de atributo (A_1, A_2, A_3,...)

Atributo		A1	A2	A3	A4	A5
Clase	Elemento					
C_1	1	1	1	0	1	1
	2	1	1	0	0	1
	3	1	1	0	1	0
C_2	4	1	0	0	0	0
	5	1	0	0	1	1
	6	1	0	0	0	1
	7	1	0	0	1	0
	8	1	0	0	0	0
C_3	9	0	1	0	1	1
	10	0	1	0	0	1
	11	0	1	0	1	0
	12	0	1	0	0	0
	13	0	1	0	1	1
C_4	14	1	0	1	0	1
	15	1	0	1	1	0
	16	1	0	1	0	0
	17	1	0	1	1	1
C_5	18	0	1	0	0	1
	19	0	1	0	1	0
	20	0	1	0	0	0

La entropía del sistema de clasificación, H(C), es:

$$H(C) = -K\sum_{i=1}^{N} p_i \log_2 p_i ; p_i = p(C_i) \tag{3.4}$$

Donde $p_i = p(C_i)$ corresponde a la probabilidad de que un elemento pertenezca a la clase C_i, $K = 1$ y N = número total de elementos. Ahora en la tabla 8 se desarrolla la ecuación anterior para obtener el valor de H (C).

Tabla 3.2 Cálculo de probabilidades de que los elementos pertenezcan a una clase y cálculo de su entropía.

Clase	Elementos	p_i	$p_i \log_2 p_i$
C_1	3	0.15	0.411
C_2	5	0.25	0.500
C_3	5	0.25	0.500
C_4	4	0.2	0.464
C_5	3	0.15	0.411
∑	20	1	**2.285**

$$H(C) = 2.285 \; bits$$

La Entropía de Shannon puede interpretarse en este ejemplo como el grado de error o incertidumbre de un problema de clasificación. En la clasificación de bialelos, la entropía de Shannon corresponde al error que se comete al asignar un bialelo aleatoriamente a una clase. Descrito de otro modo, la entropía de Shannon puede definirse como la información que se requiere obtener para resolver el problema de clasificación. Sin embargo, no dice como obtener esta información del análisis de los atributos de los elementos [5].

3.2 GANANCIA DE LA INFORMACIÓN

De acuerdo a la tesis de Rubén Fernández [5], la ganancia de la información es la herramienta que permite cuantificar la información proporcionada por un atributo *(A_K)* y permite resolver el problema de clasificación. Se define como la información transmitida por el atributo acerca de cada clase; es decir la diferencia entre la entropía de Shannon *(H(C)* y la entropía después de conocer el valor del atributo *(H (C| A_k))* [5] (ver figura 3.1):

$$I_G = H(C) - H(C \mid A_k) \tag{3.5}$$

Donde $I_G \geq 0$

También se puede definir una ganancia de la información relativa:

$$I_G R = \frac{I_G}{H(C)} \tag{3.6}$$

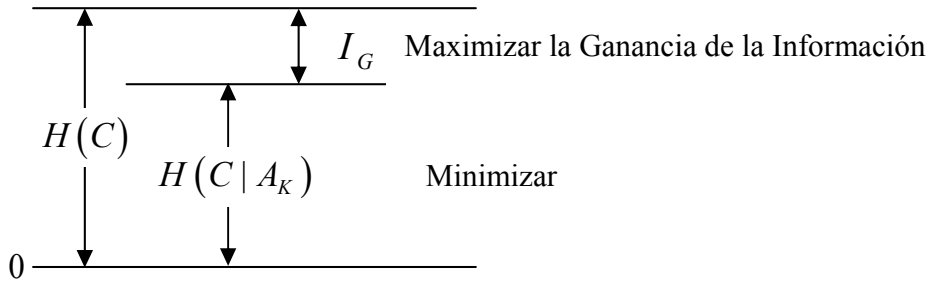

Figura 3.1: Ganancia de Información. Descripción gráfica de la ganancia de información siendo ésta la diferencia entre la entropía de Shannon antes y después de conocer el valor de un atributo.

Heurística de Selección Atributos

El atributo A_k subdivide los elementos en v_k subgrupos $S_j (j = 1,...,v_k)$ donde v_k es el número de valores que puede presentar el atributo (0, 1). Si W_i es la proporción de elementos en S_j, entonces la entropía de clasificación $H(C \mid A_k)$ después de elegir el atributo A_k, se calcula como la medida ponderada de la entropía de Shannon en cada subgrupo [5]:

$$H(C \mid A_k) = \sum_{j=1}^{v_k} W_j \times H(C \mid S_j) \tag{3.7}$$

Donde $H(C \mid S_j)$ se define como:

$$H(C \mid S_j) = -\sum_{i=1}^{N} p(C_i \mid S_j) \log_2 p(C_i \mid S_j) \tag{3.8}$$

La función $p(C_i \mid S_j)$ es la probabilidad que un elemento E_n pertenezca a la clase C_i si el elemento pertenece al subgrupo S_j. Ahora se van a desarrollar las formulas anteriores para el ejemplo de la tabla 3.1.

Para el atributo A_1 se tienen los subgrupos $S_o = \{E_9, E_{10}, E_{11}, E_{12}, E_{13}, E_{18}, E_{19}, E_{20}\}$, $S_1 = \{E_1, E_2, E_3, E_4, E_5, E_6 E_7, E_8, E_{14}, E_{15}, E_{16}, E_{17}\}$ y los valores para W_i son $W_0 = 8/20$, $W_1 = 12/20$.

El siguiente paso es obtener el valor de $H(C \mid S_j)$ para el subgrupo S_0, con la función $p(C_i \mid S_j)$ correspondiente a este subgrupo:

$$p(C_i \mid S_o) = \begin{Bmatrix} 0 & , para \quad i = 1 \\ 0 & , para \quad i = 2 \\ 5/8 & , para \quad i = 3 \\ 0 & , para \quad i = 4 \\ 3/8 & , para \quad i = 5 \end{Bmatrix} \tag{3.9}$$

$$H(C \mid S_0) = -\sum_{i=1}^{5} [p(C_i \mid S_0) \log_2 p(C_i \mid S_0)] = 0.954 \tag{3.10}$$

De igual forma se obtienen los valores anteriores para el subgrupo S_1:

$$p(C_i \mid S_1) = \begin{Bmatrix} 3/12 & , para\ i = 1 \\ 5/12 & , para\ i = 2 \\ 0 & , para\ i = 3 \\ 4/12 & , para\ i = 4 \\ 0 & , para\ i = 5 \end{Bmatrix} \tag{3.11}$$

$$H(C \mid S_1) = -\sum_{i=1}^{5} [p(c_i \mid S_1) \log_2 p(c_i \mid S_1)] = 1.554 \quad (3.12)$$

Y el valor de la entropía de clasificación para el atributo A_1, $H(C \mid A_1)$ es:

$$H(C \mid A_1) = W_0 H(C \mid S_0) + W_1 H(C \mid S_1) = 1.315 \quad (3.13)$$

La ganancia de la información para este atributo $I_G = H(C) - H(C \mid A_1) = 0.971$ con $I_G R = 42.48\%$.

En la tabla 3.3 se encuentran los valores correspondientes para los cinco atributos del ejemplo de la tabla 3.1.

Tabla 3.3 Resumen de los valores calculados para la entropía de Shannon, ganancia de información e información relativa para cada uno de los atributos

Atributo	A_1	A_2	A_3	A_4	A_5
$H(C\|A_K)$	1.315	1.293	1.564	2.246	2.246
I_G	0.971	0.993	0.722	0.039	0.039
$I_G R$	42.48%	43.44%	31.59%	1.71%	1.71%

Como se puede observar en la tabla 3.3 los atributos generan diferentes cantidades de información, el atributo A_2 genera una mayor ganancia con el 43.44%. La ganancia de información permite decidir cuál o cuáles atributos son los más adecuados para resolver un problema de clasificación en general. El objetivo es obtener una ganancia de la información I_G igual al valor de la entropía *(H(C)* de este sistema de clasificación, añadiendo los atributos necesarios y obtener una combinación única de 0 y 1 para poder identificar cada clase específicamente.