CAPÍTULO 4

HEURÍSTICA ENTROPÍA DE SHANNON Y GANANCIA DE LA INFORMACIÓN

En este capítulo se define y se plantea la metodología que sigue la Entropía de Shannon y la Ganancia de la Información para formar la heurística que será aplicada al problema de la clasificación de los alelos HLA con el fin de obtener una solución inicial de dicho problema.

4.1 CLAUDE ELWOOD SHANNON

En el año de 1916 nace Claude Elwood Shannon en Petoskey, Michigan (Estados Unidos). Estudiante destacado por su habilidad de crear prototipos técnicos y por su interés a la investigación. Graduado de la Universidad de Michigan en Ingeniería Eléctrica y en Matemáticas con premio extraordinario.

A los 24 años realizó su tesis doctoral basada en la aplicación del algebra “booleana” en el análisis de datos (An Algebra for Theoretical Genetics) en el Instituto de Tecnología de Massachussets (MIT por sus siglas en inglés) en donde también colaboró en el desarrollo de los primeros ordenadores. Su siguiente publicación la realiza a los 25 años (MathematicaL Theory of the Differential Analyzer).

Pero su principal trabajo lo publica en el año de 1948 llamado “A Mathematical Theory of Communication” en el *Bell System Technical Journal*.

Muere en Medford Massachussets en el 2001 dejando una gran contribución a la ciencia y tecnología de la comunicación de la actualidad [I2006].

4.2 TEORÍA DE LA INFORMACIÓN

La publicación “A Mathematical Theory of Communication” o también conocida como “Teoría de la Información” de C.E. Shannon crea un modelo matemático para poder descifrar los sistemas de comunicación por medio de entidades de probabilidad [P2004].

“El concepto de información es definido en términos estrictamente estadísticos, bajo el supuesto que puede ser tratado de manera semejante a como son tratadas las cantidades físicas como la masa y la energía” tal como lo menciona López et al. [LPS1995].

“Uno de los postulados básicos de la Teoría de la Información es que la ‘información’ se puede tratar como una cantidad física mesurable, tal como la densidad o la masa” [LDU2006].

La Teoría de la Información es la base en la cual se ha creado toda la teoría actual de la codificación y comunicación. Su fin es establecer los límites de cuánto se puede llegar a comprimir la información y definir la mayor velocidad en la que se puede transmitir dicha información.

La teoría se basa principalmente en el uso de la función logarítmica como una medida de información, y respalda el uso de esta función por medio de las siguientes razones [S1948]:

- Es prácticamente más usado. Los parámetros utilizados en la ingeniería tal como el tiempo, ancho de banda, el número de réplicas, etc. tienden a variar linealmente con el logaritmo.
- Se acerca más a la sensación intuitiva de la medición que se maneja en la información. Esto es aproximadamente relacionado al punto anterior desde que mide entidades por comparaciones lineales con estándares comunes.
- Es más conveniente. Muchas de las operaciones limitadas son simples, en términos del logaritmo.

La base logarítmica corresponde a la opción de una unidad para la medida de información. Por lo que el trabajo de Shannon se desarrolla en base al problema de la transmisión eficiente de la información [P2004] [S1948].

Dado que la información es tratada como magnitud física y para lograr la caracterización de una secuencia de símbolos es necesario utilizar la *Entropía*. De aquí surge lo que recibe el nombre de la “Entropía de Shannon”.

4.3 ENTROPÍA DE SHANNON

El término de entropía (del griego tropos= cambio, transformación) lo utiliza Rudolf Clausius en 1851, quien es uno de los formuladotes de la segunda ley de la termodinámica. La entropía debe ser definida tomando en cuenta consideraciones estadísticas y probabilísticas [BBML2006], una definición general de la entropía se define como "una medida de incertidumbre promedio, la cual se calcula a partir de la probabilidad de ocurrencia de cada uno de los eventos" [GHSH2002], otra definición es "Magnitud termodinámica que mide la parte de la energía que no puede utilizarse para producir un trabajo. Medida del desorden de un sistema" [W2005].

Para realizar la entropía es necesario el uso de la Teoría de la Probabilidad, cuya función básica es manejar una gran cantidad de acontecimientos o eventos que individualmente son casuales, pero en conjunto predecibles en términos probabilísticos.

El sistema total se forma por un conjunto de acontecimientos o eventos los cuales ocurren uno a la vez, es decir, "tal que uno y sólo uno de estos puede ocurrir en cada ensayo" [P2004]. Este conjunto que conforma el sistema total puede contener un número ***n*** de eventos $\boldsymbol{S}= \{E_1, E_2, E_3, E_4, \ldots, E_n\}$ cada uno de éstos con una probabilidad de ocurrencia $p_1, p_2, p_3, \ldots, p_n$ respectivamente, siempre y cuando:

$$\sum_{i=0}^{n} p_i = 1 \quad \text{donde} \quad p_i \geq 0 \qquad (4.3.1)$$

Al tener ciertos eventos con sus probabilidades de ocurrencia se tiene un "esquema finito" el cual es descrito como un estado de incertidumbre [P2004].

$$S=\begin{pmatrix} E_1 & E_2 & E_3 \cdots & E_n \\ p_1 & p_2 & p_3 \cdots & p_n \end{pmatrix}$$

Fig 4.3.1 Sistema ***S*** formado del conjunto de eventos con sus respectivas probabilidades

El estado de incertidumbre puede variar, ya que la probabilidad de que ocurra un evento en específico posiblemente sea diferente o tal vez igual a otro evento, esto es, si se cuentan con un sistema $\boldsymbol{S_i}$= {E_1, E_2, E_3) que contiene 3 eventos, las probabilidades de que ocurran se muestran en la Fig. 4.3.2. En el sistema $\boldsymbol{S_1}$ la incertidumbre es más grande ya que los tres eventos tienen la misma probabilidad de ocurrir lo cual no proporciona información útil para definir que evento puede ocurrir primero, pero en el sistema $\boldsymbol{S_2}$ se observa claramente que hay menos incertidumbre debido a que se puede decir que hay más posibilidad de que ocurra el evento $\boldsymbol{E_2}$ al primer ensayo.

$$S_1=\begin{pmatrix} E_1 & E_2 & E_3 \\ 0.333 & 0.333 & 0.333 \end{pmatrix}$$

a) Probabilidades iguales

$$S_2=\begin{pmatrix} E_1 & E_2 & E_3 \\ 0.1 & 0.5 & 0.4 \end{pmatrix}$$

b) Probabilidades diferentes

Fig 4.3.2 Probabilidades para los eventos en diferentes sistemas

Para la entropía de Shannon se maneja una alta incertidumbre ya que se tiene un sistema con dos eventos ***S* = {Si, No}** cada uno con una probabilidad del **50%** lo que forma lo que se llama **bit** ("un bit de información es suficiente para responder Verdadero/Falso a una pregunta cuya respuesta no se sabe" [CCC2006]) y por lo tanto se maneja el logaritmo en base 2. En la Tabla 4.3.1 se muestran las unidades que la entropía maneja según la base logarítmica que se esté utilizando [DIST2006].

Base Logarítmica	Unidad
2 (Binaria)	Bits
10 (Decimal)	Dits
Número e	Nats

Tabla 4.3.1Unidades para las bases logarítmicas

Para poder medir la incertidumbre que proporcionan ciertos eventos es necesario el uso de una fórmula que represente esta incertidumbre, ésta esta dada por Shannon en la ecuación de la entropía (4.3.2) la cual parte de la definición siguiente:

Definición: Se tiene un sistema de eventos $E = \{e_1, e_2, \ldots, e_i\}$ con una distribución de probabilidades dada para cada evento del sistema, dígase $P(e_i)$ se tiene que:

$$H(E) = -\sum_{\substack{e \in E \\ e \in L_0}} P(e) \times \log_b P(e) \qquad (4.3.2)$$

En donde el subíndice *b* representa la base logarítmica en la que se trabaja, en este caso la base $b=2$.

Como se observa en la Fig. 4.3.3 la incertidumbre máxima se encuentra cuando la probabilidad del evento es de 0.5, contrario a los valores de la probabilidad 0 y 1 en

donde la entropía tiene un valor de cero ya que se conoce definitivamente si el evento se realiza o no se realiza, por lo cual no hay incertidumbre alguna [CCC2006] [DIST2006], y esto representa una ganancia de información, que más adelante se explicará con detalle.

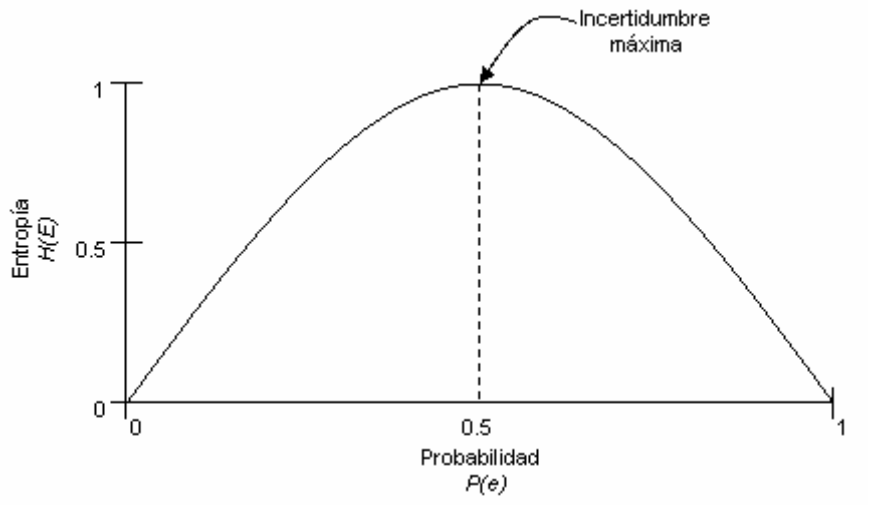

Fig. 4.3.3 Representación de la Función de Entropía

Cabe mencionar que la entropía de Shannon también se basa en la información que un atributo puede proporcionar. “Cada aspecto que se representa sobre una entidad o evento se denomina *atributo*” [J2006].

4.4 GANANCIA DE LA INFORMACIÓN

La Ganancia de Información (I_G) se define como “la reducción de la entropía causada por particionar un conjunto de eventos *S*, con respecto a un atributo A_k” tal como lo menciona Guerra [G2004]. Esta ganancia de información “es la herramienta que permite cuantificar la información proporcionada por un atributo A_k” [F2002].

La ganancia de información se obtiene de la resta entre la entropía de Shannon de todo el sistema (H(E)) y la entropía obtenida de la información que proporciona el atributo correspondiente (H(E|A_k)), en donde (E|A_k) es la ocurrencia de E dado un atributo [P2004]. En la ecuación 4.4.1 se muestra la fórmula de la ganancia de información. Pero la ganancia de información también puede ser representada de manera porcentual obteniéndose de la ecuación (4.4.2) la Ganancia de Información Relativa (I_GR).

$$I_G = H(E) - H(E \mid \mathrm{A_k}) \quad \text{donde} \quad I_G \geq 0 \qquad (4.4.1)$$

$$I_G R = \frac{I_G}{H(E)} \qquad (4.4.2)$$

Así que, como lo muestra Peña [P2004] en la Fig. 4.4.1, el objetivo es maximizar la ganancia de información que se pueda obtener de un atributo y minimizar la entropía que éste genera. Al minimizar la entropía (H(E|A_K)) se evita tener incertidumbre sobre el atributo, como se mencionó anteriormente.

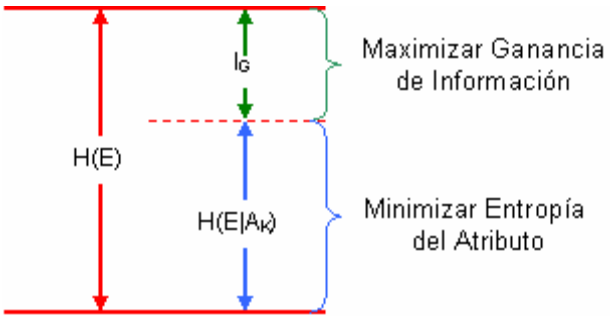

Fig. 4.4.1 Obtención de la ganancia de información. Fuente Peña [P2004]

Con esta información acerca de la entropía de Shannon y la ganancia de información es posible realizar una heurística para la clasificación de alelos y la cual es descrita en el siguiente capítulo.