DOI: 10.30972/eitt.704764

Teoría de la información y codificación: algunas consideraciones sobre la codificación al azar en Canal Binario Simétrico (BSC) con ruido

Magister Ingeniero Emilio Fabián Scozzina(*)

Resumen

Cuando Claude E. Shannon, expuso su trabajo *"A Mathematical Theory of Communication (1948)"* sentó las bases de las matemáticas de la teoría de la información y las comunicaciones modernas, dio los fundamentos matemáticos para industrialización de la información, tan vigente en nuestros días. En esta publicación, presentaremos algunos aspectos que de describen la codificación al azar, como base de la teoría de codificación en canales ruidosos. Describiremos algunas consideraciones del teorema de codificación con ruido, o teorema fundamental de Shannon para un canal con ruido; Sabemos que la capacidad de un canal, es una propiedad fundamental, y esto cobra singular importancia cuando tratamos de transmitir información sobre un canal con ruido, a una velocidad menor que dicha capacidad, con una probabilidad de error lo suficientemente aceptable. Los resultados, son la base que han permitido desarrollar numerosas técnicas de codificación, aplicables a sistema de comunicaciones.

Palabras Claves: Codificación, Canal Binario Simétrico, Ruido, Shannon

1. Introducción.

Para el desarrollo consideramos una fuente de información de memoria nula (FMIN), que llamamos $\{A_m\}$, a la salida de esta fuente se emite un alfabeto de a_m

(*) Contacto: efscozzina@gmail.com; efscozzina@exa.unne.edu.ar | Tel. Celular: 0362-15-4527366

(símbolos fuente), todos ellos equiprobables e independientes, la entropía de esta fuente $\{A_m\}$, estará dada por la formula clásica:

$$A_m = \{a_1, a_2, a_3, \dots a_k .. a_m\}$$

$$H(A) = \sum_{i=1}^{M} P(m)_i \, log_2 \frac{1}{P(m)_i} = log_2 m$$

Como segundo paso, si a la fuente $\{A_m\}$, le adicionamos a su salida un codificador, de tal manera que podamos codificar el alfabeto de la fuente original {Am}, con un alfabeto código que llamaremos $\{D_m\}$. A condición que el alfabeto código $\{D_m\}$ deba tener ciertas propiedades, como ser que la longitud (l_{Di}=n) de cada una de las palabras códigos binarias será de n bits, es decir codificamos el alfabeto original {Am}, con una secuencia muy larga de n símbolos.

Puede surgir una pregunta, como seleccionamos las palabras? una importante es que tomemos al azar todas las palabras $\{D_m\}$, de un conjunto donde m sea menor o igual a 2^n, valores posibles. Destacamos que la probabilidad de ocurrencia de estos símbolos, será $p(m/2^n)$, y también son equiprobables e independientes.

$$D_m = \{d_1, d_2, d_3, \dots d_k .. d_m\}, todos\ de\ longitud = n$$

Entropía de entrada al canal por digito será:

$$H(D) = \sum_{i=1}^{M} P(d)_i \, log_2 \frac{1}{P(d)_i} = log_2 m$$

La longitud será:

$$l = \sum_{i=1}^{M} P(d)_i \, l_i$$

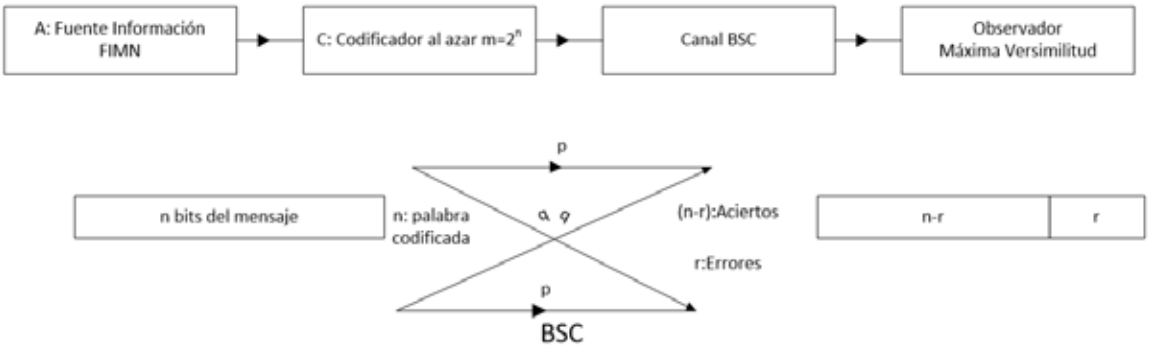


Figura: *Modelo de fuente, codificador y BSC*

Ahora, agregamos un BSC al final de codificador. Las palabras luego de ser codificadas, son transmitidas por un canal BSC, de memoria nula. El BSC, cuenta con las siguientes probabilidades, p es la probabilidad de que los datos lleguen sin corrupción (sin error) y q es la probabilidad de error (corruptos), siempre se verifica que $(p + q)=1$.

Se debe cumplir por regla general, que q sea menor a 1/2, para que el canal sea operativo, la capacidad del canal está dada por la ecuación:

$$C = 1 + (p \times log_2 p + q \times log_2 q)$$

El decodificador es un observador ideal, que reanaliza la función de máxima verosimilitud. El canal BSC se comporta como un proceso de Bernulli, donde sobre n pruebas, existen (n-r) probabilidades de éxito y r fracasos.—*Esto último se puede asimilar a una trama de datos, de longitud n*— Supongamos que deseamos enviar un mensaje y tomamos al azar una palabra, del conjunto de 2_n, posibles. Al pasar por el canal, debido al ruido, la palabra se altera en r símbolos, mientras los otros (n-r) continúan sin alteración. La distribución binomial sigue la forma clásica:

$$p\{r\ errores\} = \binom{n}{n-r} p^{n-r} q^r = \frac{n!}{r!\,(n-r)!} p^{n-r} q^r$$

Donde r =0,1,2,3,4... el número de fracasos sobre, n pruebas. Está demostrado que en un proceso de Bernulli tiene un valor medio de errores, ó en este caso de dígitos alterados por el ruido en el canal, que vendrá dado, por:

$$E\{r\} = nq = n(1-p)$$

De manera sencilla podemos calcular en cuantas secuencias binarias de longitud n pueden diferir de las palabras enviada a través del BSC, (si existen 2_n palabras posibles del código). La fórmula de K encierra un concepto muy rico, que representa el conjunto de palabras que pueden diferir en (nq) dígitos o menos, de la palabra originalmente transmitida en el mensaje a través del BSC.

$$K = \sum_{r=0}^{nq} \binom{n}{r} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} \dots + \binom{n}{nq-1} + \binom{n}{nq}$$

Por ejemplo, consideramos una palabra de tres dígitos, n=3 y $2^n=2^3=8$. Supongamos que tomamos una palabra al azar, por ejemplo (000), hay tres palabras con un error (001,010,100) y tres palabras en dos errores (011,101,110) y una con tres errores (111) Hay que comprender, que para un detector es difícil discernir que palabras fueron enviadas, en el caso de aparezcan dos errores, ya no podría distinguir si la palabra enviada es (000) ó (111). Un detector ideal realizaría la máxima verisimilitud en la secuencia.

$$K = \sum_{r=1}^{nq} \binom{n}{r} = \binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 1 + 3 + 3 = 7$$

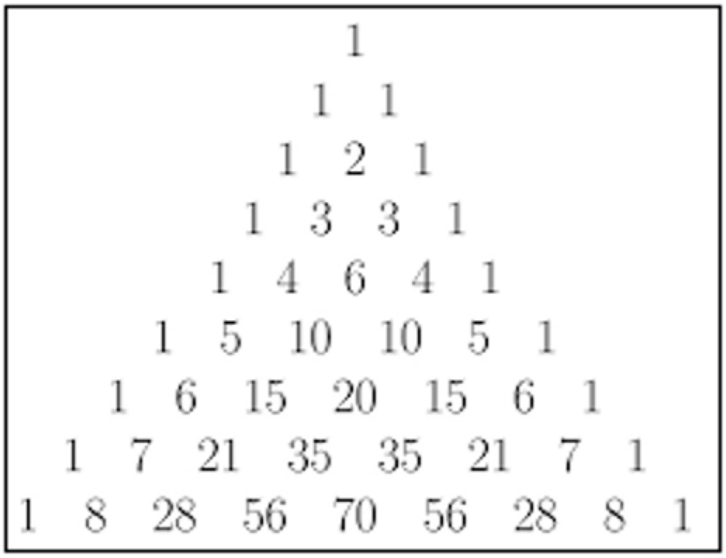
O el caso de una palabra de longitud n=8 y nq hasta 3 errores.

$$p^8 + 8p^7q^1 + 28\,p^6q^2 + 56\,p^5q^3 + 70\,p^4q^4 + 56\,p^3q^5 + 28\,p^2q^6 + 8\,p^1q^7 + q^8$$

$$K = \sum_{r=1}^{nq} \binom{n}{r} = \binom{8}{1} + \binom{8}{2} + \binom{8}{3} = 8 + 28 + 56 = 92\ diferentes\ de\ 256$$

Dejamos al lector puede probar otras relaciones, recordando de que el binomio de Newton se expresa según:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$



Existen distintos trabajos, que tratan de buscar una acotación de la serie K para logar un resultado general y obtener conclusiones sobre este tipo de codificación.

Pero para consideraciones prácticas, el valor de *q*, siempre es menor 0,5 y la expresión queda acotada, para secuencias de codificación muy largas de n bits. Esto implica que al menos dígitos n/2=r pueden ser erróneos (la mitad de la trama). La expresión representa el conjunto de palabras de K que difieren de la palabra enviada, la serie se encuentra acotada por el producto de (*nq+1*), es decir el valor medio de dígitos de error, más un dígito.

$$K \leq (nq+1)\binom{n}{nq} = (nq+1)\left(\frac{n!}{(nq)!\,(n-nq)!}\right)$$

Aplicando la fórmula de Stirling para aproximación del cálculo factorial de un número, trataremos de llegar a una expresión más compacta.

$$n! \approx n^n \sqrt[2]{2n\pi}\, e^{-n} \approx \sqrt[2]{2\pi}\, n^{\frac{n+1}{2}} e^{-n}$$

$$K \leq \sqrt[2]{\frac{(nq+1)^2}{2\pi nqp}}\, q^{-nq} p^{-nq}$$

Una palabra enviada, puede ser recibida como cualquiera de las otras palabras código, que difieren de la palabra enviada en *nq* dígitos o menos. Recordando la condición inicial, que las *m* palabras código fueron asignadas al azar, de un conjunto de 2^n palabras posibles, entonces el número esperado de mensajes L que atraviesan el BSC y pueden confundirse con el mensaje original se puede expresar como:

$$L = K\left(\frac{m}{2^n}\right) = \left[\frac{palabras\ con\ error\ \{K\} \times\ mensajes\ m\ al\ azar}{Posibles\ palabras\ codigo\ 2^n}\right]$$

Por ejemplo, si seleccionamos *m* = 32 símbolos, tomados al azar de un alfabeto de 216=65.536 palabras posibles, con una longitud *n* de 16 bits, cada símbolo, esto nos da una probabilidad bastante pequeña, multiplicada por el conjunto de palabras K. L representa el número de mensajes, y está acotado:

$$L \leq m\left(\sqrt[2]{\frac{(nq+1)^2}{2\pi npq}}\right)2^{-nC}$$

Para palabras largas, es decir con una secuencia de dígitos binarios muy grande, L tiende a cero, relacionamos el número de mensajes *m*, con la capacidad del canal C, reemplazando en L.

$$m = \left(\frac{2^{nC}}{n}\right)$$

$$L \leq \sqrt[2]{\frac{(nq+1)^2}{2\pi n^3 pq}}$$

Físicamente esto nos indica que, para valores convencionales de la capacidad de un BSC, y tomado un conjunto de *m* símbolos, si la longitud *n* tiende a infinito, la cantidad de palabras recibidas con errores tiende a cero. (Que se decodifique erróneamente un mensaje).

$$limite_{n\to\infty} L = 0$$

La conclusión más interesante es que la información que puede transmitirse por un canal BSC tiende a la capacidad del canal, para secuencias para secuencias muy largas de *n*. La entropía por digito será H(D)`:

$$H(D)` = \left(\frac{H(D)}{n}\right) = \frac{log_2 m}{n}\ \left[\frac{bits}{digitos}\right]$$

$$H(D)lim_{n\to\infty} = \left[\frac{log_2 \frac{2^{nC}}{n}}{n}\right] = C - \left(\frac{log_2 n}{n}\right) = C\ \left[\frac{bits}{digitos}\right]$$

Se puede demostrar que, con secuencias n lo suficientemente largas, la velocidad de información de tiende a la capacidad del canal, reduciendo la probabilidad de errores. Dicho de otra manera: Si consideramos un canal BSC, sin memoria y con una capacidad no nula, y fijamos dos números positivos, mayores que cero H y ε, tal que:

$$0 < H < C$$

$$\epsilon < 0$$

De tal modo que transmitimos *m* palabras de longitud *n*, que cumplan con la condición:

$$m \leq 2^{nH}$$

Donde la probabilidad de error no exceda ε. Se comprueba la siguiente regla de decisión dada por la probabilidad condicional:

$$p\left[{}^{d_i}\!/_{d_j}\right] \geq (1-\varepsilon)$$

Por los cual puede realizarse una decodificación con una probabilidad condicional de error que no exceda ε.

Conclusiones:

Existe una demostración que es la inversa, la cual expone que: Es imposible producir un procedimiento de codificación que permita la transmisión sobre un BSC con ruido a velocidades superiores a la capacidad del canal.

$$m \geq 2^n$$

Recordemos que para un número secuencia binarias n, lo suficientemente grande, se puede formar un subconjunto de m palabras equiprobables que están relacionadas con las 2^n entradas del $(BSC)^n$. Es lógico pensar que la probabilidad de error aumentara si las palabras están seguidas (secuencia continua) unas otras y no existen una distancia de Hamning d:(di;dj) adecuada, es decir que el procedimiento de codificación tiene singular importancia en la probabilidad de error y el número máximo de palabras código que pueden utilizarse.

Si bien en la explicación, existe una simplificación, debemos considerar que para el procedimiento de codificación al azar, puede considerarse como el caso, introducir 2^n bolillas con el código binario estampado en ellas, en una tómbola, y luego de girarla, sacar un al azar y luego regresar (reponer) la bolilla a la tómbola, con los cual existen 2^{mn} palabras códigos, y pueden existir palabras repetidas, esta es una limitación, y no se ocupa en la práctica.

El teorema no dice como construir el código, pero sus resultados han abierto un camino para los numerosos sistemas de codificación que actualmente están en uso.

Siempre los códigos se idean para tener una longitud mínima, próxima al valor de entropía de la fuente, en el caso de transmisión sobre un canal sin ruido, en los canales ruidosos es necesario introducir redundancia. El segundo teorema de Shannon para un canal con ruido, simplemente nos indica que es posible transmitir fiablemente con secuencias largas y a una velocidad muy próxima a la capacidad del canal.

Bibliografía:

Thomas M. Cover, Joy A. Thomas,*"Elements of Information Theory"*, John Wiley & Sons. Second Edition 2006

Zanuy, Fernadez, Marcos Tratamiento digital de voz e imagen y su aplicación a la multimedia. Alfaomega Marcombo 2001

Skar, Bernad "*Digital Communications Fundamentals and applications"*. Segunda edición en ingles.Prentice Hall PTR 2000

Norman Abramson, "Teoria de la Información y Codifiación" Paraninfo Madrid Tercera edición. 1974 Edición original en inglés 1966.

Claude E. Shannon. *"A Mathematical Theory of Communication" published in Bell System Technical Journal in 1948.*

Donald Fink, Alexander McKenzie, *Electronic Engenier `s Handbook Mc Graw Hill Frist Edition 1975.*

Digital Communications Fundamentals and Applications. Bernard Sklar. 2ª Edition. Prentice Hall PTR.