Universidad Nacional de Mar del Plata.
Facultad de Ingeniería.
Departamento de Ingeniería informática.
Asignatura: Teoría de la Información.

Trabajo Practico Integrador

Informe

2° parte

Grupo N°3

Integrantes:

- Cacace, Camila (16403)
- Gentili, David (12940)
- Luna, Lautaro (16175)

Repositorio: https://github.com/DavidGentili/teoria-de-la-informacion-tp2

Video: https://www.youtube.com/watch?v=jOV5P2R9iS0

Fecha de entrega: 22/11/2022

Índice

Resumen

En este segundo trabajo integrador se busca, mediante un texto provisto por la cátedra, aplicar algoritmos de codificación y decodificación, para luego cargarlas en un archivo con su extensión correspondiente.

Introducción

Antes de comenzar a desarrollar los cálculos y algoritmos implementados para el trabajo integrador daremos una breve introducción de manera que se entienda con mayor facilidad el marco teórico con el que elaboramos el código.

Compresión de datos

Uno de los temas que toca la codificación es la compresión de datos, esta permite aumentar la capacidad de almacenamiento de los dispositivos y transmitir información por un canal en menor tiempo. Existen varios métodos de compresión, divididos en **métodos sin pérdida** o **métodos con pérdida**, es decir, métodos con tasas de compresión bajas pero que mantienen la integridad de la información y métodos con tasas de compresión altas en los que los datos descomprimidos difieren a los datos originales, respectivamente. Una particularidad de los métodos sin pérdida, o *lossless compression*, es que la longitud media del código es mayor o igual a la entropía del mismo.

Los métodos de compresión utilizados en este trabajo fueron los indicados por la cátedra: Shannon-Fano y Huffman. Estos pertenecen a los métodos sin pérdida, y

Rendimiento y redundancia

Gracias al primer teorema de Shannon se pudo demostrar que el valor medio de un símbolo de S es H (S). De forma más general, el valor medio de un símbolo de S, en dígitos r-arios, es $H_r(S)$. Supongamos que L es la longitud media de un código r-ario, unívoco, de la fuente S. L no puede ser inferior a $H_r(S)$.

Según esto, se define η , **rendimiento del código**, como :

$$\eta = \frac{H_r(S)}{L}$$

Igualmente, puede definirse la redundancia de un código como:

$$Redundancia = 1 - \eta = \frac{L - H_r(S)}{L}$$

(Abramson, 1986, Pág. 102).

Canales de comunicación

A partir de la codificación podemos mencionar un nuevo término que es el de los canales de información. Se trata de medios mediante los cuales se transmite información desde la fuente hasta el destino. Esta información que viaja es codificada antes de entrar al canal y decodificada cuando sale.

Estos canales, vienen determinados por un alfabeto de entrada A = {ai}, i = 1, 2, ..., r y un alfabeto de salida B = {bj}, j = 1, 2, ..., s, acompañados de un conjunto de probabilidades condicionales al que nos referiremos como P(bj/ai), es decir, la probabilidad de recibir la salida bj habiendo enviado una entrada ai.

Los canales de información están definidos completamente por su matriz, las cuales representan en sus filas el alfabeto de entrada, en sus columnas el alfabeto de salida y en su interior todas las probabilidades condicionales que correspondan con la entrada y salida de los símbolos. Esto cumpliendose siempre una condicion fundamental que es la siguiente:

$$\sum_{j=1}^{s} Pij = 1 \, para \, i = 1..r$$

Relaciones entre probabilidades de un canal

Estas matrices presentan también otras relaciones entre sus probabilidades cuyo uso en diferentes cálculos puede aportar información valiosa acerca de los canales bajo estudio, por ejemplo el cálculo de la probabilidad independiente P(bj):

$$P(bj) = \sum_{i=1}^{r} P(ai) \, Pij \;\; \forall j = 1.. \, s$$

El cálculo de la probabilidad condicional P(ai/bj), es decir que se haya enviado el símbolo ai, habiendo recibido el símbolo bj:

$$P(ai/\, bj) = \frac{p\,(bj/\, ai) P(ai)}{P(bj\,)}$$

Y la probabilidad de que ocurra un suceso simultáneo:

$$P(ai,\, bj) = P(ai/bj)\, P(bj) = P(bj/ai).\, P(ai)$$

A partir de estas relaciones, podemos rescatar diferentes tipos de datos que nos ayudarán a entender nuestros canales de información. Entre ellos podemos destacar los cálculos de las entropías, tanto “a-priori” como “a-posteriori”.

Como ya sabemos, el valor de la entropía simboliza una medida de la incertidumbre sobre el símbolo que emitirá una fuente. En el caso de la entropía “a-priori”, nos referimos a un caso previo a recibir un símbolo de salida, por ende hacemos referencia solamente a la probabilidad de entrada del símbolo P(ai). Su cálculo sería el siguiente:

$$H\,(\,A) = \sum_{A} P\,(a)\, log\, \frac{1}{P(a)}$$

En cambio, para el cálculo de la entropía “a-posteriori”, tomamos en cuenta la salida obtenida bj. Por lo que para su cálculo precisamos la probabilidad condicional P(ai/bj), quedándonos entonces:

$$H\,(\,A\,/\,bj) \;=\; \sum_{A} P\,(a\,/\,bj)\,log\frac{1}{P(\,a\,/\,bj)}$$

En ambos casos, basándonos en el primer teorema de Shannon, que especifica cómo se ha de interpretar la entropía, lo que representan estos cálculos es el número medio de binits necesarios para representar un símbolo de una fuente con la probabilidad correspondiente.

Equivocación de un canal

El valor medio de las entropías a posteriori también recibe el nombre de equivocación A con respecto a B. Este valor se calcula como:

$$H\,(\,A\,/\,B) \;=\; \sum_{B} P\,(b)\,H\,(\,A\,/b)$$

La equivocación de un canal representa la información que queda en A después de observar B, la pérdida de información sobre A causada por el canal mismo, es decir, la cantidad de información sobre A que no deja pasar el canal.

Información mutua

Según el primer teorema de Shannon la determinación de un símbolo de entrada ai exige una media de H(A) binits. En el caso de conocer el símbolo de salida producido por una entrada, la observación de un símbolo de salida proporciona H(A/B) binits para definirlo. Entonces, definimos a la diferencia H(A) - H(A/B) como **información mutua**, que representa la cantidad de bits de información que proporciona la información de un símbolo de salida. Esta se escribe:

$$I(A,B) \;=\; H(A) \;-\; H(A/B)$$

Es la cantidad de información sobre A, menos la cantidad de información que todavía hay en A después de observar la salida.

Es un indicador de la información ganada debido al acople entre las variables A (entrada) y B (salida).

Desarrollo

Primera Parte: Codificación y compresión

Recolección de datos

Para poder determinar el alfabeto recorrimos todo el archivo, almacenando todas las palabras, sin repetirlas, y contando su cantidad de apariciones, además de la cantidad total de palabras. Cabe aclarar que, por convención de la cátedra, las palabras acompañadas de signos de puntuación fueron tomadas como palabras nuevas y no como una aparición de la palabra sin tales signos. En total se encontraron 4150 palabras distintas.

Luego, la probabilidad de cada palabra se calculó como el cociente entre su frecuencia absoluta y la cantidad total de palabras.

Una vez obtenidas las probabilidades de ocurrencia de cada una de las palabras, se implementaron los algoritmos de Huffman y Shannon-Fano, con el fin de obtener sus respectivas codificaciones.

Luego se procedió a codificar el mensaje, con dichos códigos, generando así los mensajes codificados. Por último, se almaceno la tabla y el mensaje codificado en un archivo utilizando la siguiente estructura:

Primero la tabla que comienza con un entero que determina la longitud de la tabla en bits, luego el conjunto palabra-código almacenando la palabra, el separador '|', un byte que determina la longitud en bits del código y dicho código. Luego de esto un entero que determina la longitud del mensaje codificado en bits, y el mensaje codificado en binario.

Luego de la creación de dichos archivos se obtuvieron los siguientes valores:

Huffman

Longitud Media	Entropía	Rendimiento	Redundancia
9,40	9,38	0,9972	0,0028

Tamaño de la tabla	Tamaño del mensaje	Tamaño del archivo comprimido	Tamaño del archivo original	Tasa de compresión
46.394 bytes	17.000 bytes	63.394 bytes	83.138 bytes	1,3114

Shannon-Fano

Longitud Media	Entropía	Rendimiento	Redundancia
9,43	9,38	0,9946	0,0054

Tamaño de la tabla	Tamaño del mensaje	Tamaño del archivo comprimido	Tamaño del archivo original	Tasa de compresión
46.346 bytes	17.045 bytes	63.391 bytes	83.138 bytes	1,3115

Interpretación de los resultados

Tanto la codificación generada a partir del algoritmo de Huffman como la resultante del algoritmo de Shannon-Fano, tiene un rendimiento óptimo, resultando, la codificación de Huffman, levemente superior.

Con respecto a la compresión, se puede ver cómo el mensaje reduce su tamaño de manera representativa, pero es la tabla quien se encarga de ocupar gran parte del tamaño del archivo comprimido, representando el 73% del contenido del archivo en ambos casos.

En este caso, podemos observar como la compresión del mensaje es levemente superior con huffman, pero con la tabla incluida, es el algoritmo de Shannon.Fano el que obtiene un mejor resultado, con una tasa de compresión levemente mayor.

Segunda Parte: Canales de comunicación

Cálculo de matrices

Al calcular los valores de las celdas con condiciones puestas por la cátedra, obtuvimos las siguientes matrices para cada canal:

Canal 1

	B1	B2	B3
S1	0.3	0.3	0.4
S2	0.12	0.4	0.48
S3	0.3	0.3	0.4
S4	0.3	0.4	0.3
S5	0.3	0.12	0.58

Canal 2

	B1	B2	B3	B4
S1	0.2	0.3	0.3	0.2
S2	0.3	0.3	0.2	0.2
S3	0.3	0.2	0.2	0.3
S4	0.3	0.3	0.3	0.1

Canal 3

	B1	B2	B3	B4
S1	0.2	0.3	0.2	0.3
S2	0.3	0.3	0.3	0.1
S3	0.2	0.2	0.3	0.3
S4	0.3	0.3	0.2	0.2
S5	0.2	0.3	0.3	0.2
S6	0.2	0.3	0.3	0.2

Obtención de resultados

A partir de las matrices de los tres canales de información provistas por la cátedra, al igual que los cálculos que nos brindaron para completarla y asegurándonos que cumpliera con la condición fundamental mencionada anteriormente $\sum_{j=1}^{s} Pij = 1\, para\, i = 1..r$, logramos armar los canales con sus probabilidades condicionales correspondientes para luego poder continuar con los cálculos pedidos.

Inicialmente, debíamos encontrar los valores de las probabilidades de los símbolos de la salida. Para esto tomamos la ecuación correspondiente $P(bj) = \sum_{i=1}^{r} P(ai)\, Pij$, cuyas variables teníamos como dato y, decidimos combinarla directamente con la ecuación pertinente al cálculo de sucesos simultáneos $P(ai, bj) = P(bj/ai).\ P(ai)$, ya que estos últimos podrían ser de utilidad más adelante. Por lo tanto, al armar la matriz de sucesos simultáneos y para cada columna de la misma sumar sus valores obtuvimos tablas como las que se muestran a continuación

Canal 1

Salida	
Símbolo	**P(i)**
B1	0.28
B2	0.32
B3	0.40
Suma	1.0

Canal 2

Salida	
Símbolo	**P(i)**
B1	0.28
B2	0.27
B3	0.24
B4	0.21
Suma	1.0

Canal 3

Salida	
Símbolo	**P(i)**
B1	0.21
B2	0.28
B3	0.29
B4	0.23
Suma	1.0

Al obtener un valor aproximado a 1 en la sumatoria de las probabilidades al final de cada tabla, concluimos que los valores calculados eran correctos, ya que se cumple que la sumatoria de las probabilidades de todos los símbolos tanto, de entrada (también corroborado) como de salida fueran 1.

A partir de obtener las probabilidades de los símbolos de salida y a su vez la matriz de sucesos simultáneos continuamos con el cálculo de las probabilidades “a-posteriori”, que como mencionamos anteriormente, se trata de las probabilidades asociadas a la entrada después de la recepción del símbolo de la salida. Éstas también partieron de la de primera ecuación de sucesos simultáneos, que al reformular, nos da la información que buscamos, las probabilidades “a-posteriori” P(ai/bj).

Mediante el uso de cuentas auxiliares, para facilitar el uso de la hoja de cálculos, obtuvimos de todos los canales las entropías “a-posteriori”. Estos cálculos representan el número medio de binits necesarios para representar el símbolo de la fuente con la probabilidad correspondiente P(ai/bj). Los cálculos de estas entropías para los tres canales fueron:

Canal 1

Símbolo	H(A/bj)
B1	2.06
B2	2.05
B3	2.25

Canal 2

Símbolo	H(A/bj)
B1	1.92
B2	1.93
B3	1.97
B4	1.82

Canal 3

Símbolo	H(A/bj)
B1	2.14
B2	2.09
B3	2.07
B4	2.03

Con todo lo anterior ya calculado, pudimos comenzar a realizar los cálculos que nos interesaban para poder sacar las conclusiones necesarias para un profundo análisis de los canales.

Resultados de los canales

Para el primer cálculo pedido, el de la equivocación, utilizamos la ecuación de la entropía media “a-posteriori”. En este caso, se trata de la equivocación de A con respecto a B. Los resultados obtenidos fueron los siguientes:

	Equivocación
Canal 1	2.13
Canal 2	1.92
Canal 3	2.08

La información que nos brinda este cálculo se puede interpretar de varias maneras, que en esencia representan lo mismo. Para poder relacionarlo más adelante con otro término, vamos a decir que es la **pérdida de información sobre A causada por el canal.**

Ahora, al realizar el cálculo de la entropía “a-priori” H(A), que en este caso se refiere a el número medio de binits necesarios para representar un símbolo de una fuente con una probabilidad a priori P (ai) sin conocer la salida, conseguimos la siguiente tabla:

	H(A)
Canal 1	2.17
Canal 2	1.95
Canal 3	2.53

Con estas dos últimas tablas podemos realizar un último cálculo, que es el de la información mutua. Esta es la resta entre los resultados obtenidos anteriormente y representan la **cantidad de información que se obtiene de A gracias al conocimiento de B** o también se puede ver como la incertidumbre sobre la entrada del canal que se resuelve observando la salida del canal. Los resultados fueron:

	Información mutua (H(A) - H(A/B))
Canal 1	**0.04**
Canal 2	**0.03**

Canal 3	0.45

Interpretación de los resultados

Con los resultados previos podemos sacar algunas conclusiones para cada canal.

En el caso del canal 1, al comparar la entropía “a-priori” con cada una de las entropías “a-posteriori” correspondientes a una salida, vemos que solo en el caso del símbolo de salida B3 se cumple $H\,(A/bj) >= H\,(A)$, por lo que simplemente en este caso se puede asegurar que **hay más incertidumbre al observar la salida**.

En el canal 2, ocurre lo mismo, el símbolo B3 es el único que cumple la condición anterior por lo que, al igual que en el canal previo, cuando se recibe ese símbolo de salida se halla mayor incertidumbre que en la entrada.

En cambio, en el canal 3 no existe ninguna salida en la que suceda esto.

Para todos los canales se cumple la condición $H\,(A\,/\,B) < H\,(A)$, por lo que podemos afirmar que en promedio **nunca se pierde información al conocer la salida**.

Esto mismo nos lo confirma la información mutua, cumpliendo $I\,(A\,,\,B) \geqslant 0$ para todos los canales. Además la condición para que la información mutua sea nula es que los símbolos de entrada y salida sean estadísticamente independientes.

Conclusiones

La resolución de este trabajo mediante la aplicación de los temas tratados durante la primera parte de la cursada nos permitió poder volcar todo lo visto en dos casos de estudio. En uno de ellos aplicando codificación y decodificación, y en el otro las propiedades de los canales de información.

Al realizar la compresión de un archivo pudimos indagar sobre dos métodos de codificación y distintas formas de implementar el almacenamiento del archivo comprimido, además de entender que dichas implementaciones tendrán mejores o peores resultados dependiendo del conjunto de datos.

Al analizar en profundidad estos 3 canales de comunicación, pudimos familiarizarnos con los conceptos relacionados con la información que brinda un canal de transmisión, tal como probabilidad de sucesos simultáneos, entropías “a-priori” y “a-posteriori”, equivocación e información mutua.

Para concluir, nos gustaría destacar el hecho de que este trabajo nos ayudó a dimensionar la complejidad que tiene la compresión de datos para brindar soluciones realmente eficientes.

Referencias

Abramson, N. (1986). *Teoría de la Información y Codificación* (J. A. d. Miguel Menoyo, Trans.). Paraninfo.