Trabajo Práctico Integrador 2

Integrantes:

- de Lellis, Lucas (lucasdelellis2002@gmail.com)
- Ramos Kees, Teo (teoramites@gmail.com)
- Romeo, Renzo Agustin (renzoromeo44@gmail.com)

Índice

Resumen

El objetivo del trabajo fue comprender los métodos de compresión y analizar la transmisión de información en canales de comunicación. Para esto, primero se realizó un análisis de dos métodos de compresión diferentes: el algoritmo de Huffman y el algoritmo de Shannon-Fano. Utilizando estos algoritmos, se comprimió un archivo de texto y se analizaron la tasa de compresión, rendimiento y redundancia de los mismos.

Por otra parte, se realizó un análisis sobre tres canales de comunicación diferentes definidos por las probabilidades de los símbolos de entrada y su matriz de canal. En base a los mismos, se calcularon la equivocación, la información mutua y la entropía afín, y se compararon los resultados obtenidos para cada canal.

Introducción

En general, los archivos contienen mucha redundancia. Para evitar esto existen los algoritmos de compresión, que reducen el tamaño del archivo permitiendo ahorrar espacio de almacenamiento y tiempo al transmitirlo. Existen de dos tipos, con pérdida y sin pérdida. Los algoritmos sin pérdida mantienen la integridad de la información, es decir, permiten recuperar el archivo original a partir del archivo comprimido. Además, cumplen que la longitud media del código es mayor o igual a la entropía. Por otro lado, los algoritmos con pérdida permiten una mayor compresión ya que no mantienen la integridad de la información.

En este caso, se utilizaron dos algoritmos de compresión sin pérdida basados en diccionarios adaptativos. Este tipo de algoritmos adaptan la distribución de probabilidades de las palabras basándose en los datos a codificar. Esto permite una mayor compresión pero requieren procesar el archivo original para poder armar la tabla y para luego almacenarla en el archivo comprimido. Dos algoritmos de este tipo son el de Huffman y de Shannon-Fano.

Huffman

El algoritmo de Huffman permite generar un código binario compacto óptimo en base al código dado. Para llevar a cabo el algoritmo, se ordenan las palabras del alfabeto original en orden descendente de probabilidad. Luego, las dos palabras con menor probabilidad se agrupan y se suman sus probabilidades. Se realiza este proceso hasta obtener un alfabeto con solo dos palabras, y se les asigna un dígito binario a cada una. Luego, se realiza el proceso en reversa, agregando un dígito binario más a las dos palabras que fueron agrupadas anteriormente, hasta finalmente obtener un código para el alfabeto original. La característica principal del código generado, es que es un código compacto que asigna a las palabras más frecuentes un código de menor longitud.

Shannon-Fano

El algoritmo de Shannon-Fano permite generar un código binario compacto subóptimo en base al código dado. Para llevar a cabo el mismo, inicialmente se ordenan las palabras del alfabeto original en orden descendente de probabilidad. Luego se elige un valor entero positivo k menor o igual a la cantidad de palabras del alfabeto, de modo que la diferencia entre la suma de probabilidades de las palabras en posiciones menores o iguales a k y la suma de las probabilidades de las palabras en posiciones mayores a k sea mínima. Luego, a las palabras en posiciones menores a k se le asigna un dígito binario y al resto se le asigna el otro. Este proceso se repite recursivamente para ambos intervalos de palabras resultantes hasta llegar a grupos de dos palabras, donde se asigna '0' a una de ellas y '1' a la otra. El código final resultante es la concatenación de los dígitos binarios asignados a cada palabra en cada paso.

Tasa de Compresión

Con los códigos binarios es posible regenerar el archivo original de forma comprimida. Este archivo contiene las palabras originales en su forma binaria y una tabla que contiene las palabras originales y su correspondencia en el código binario. Con esta tabla es posible descomprimir el archivo para poder interpretarlo.

En siguiente lugar, es posible calcular la tasa de compresión de los archivos generados. Esta tasa representa el porcentaje de espacio que ocupa el archivo comprimido respecto al original y se calcula mediante la siguiente fórmula:

$$tasa\ compresión\ = \frac{peso\ comprimido}{peso\ original}$$

Fórmula 1: Tasa de compresión.

Rendimiento y Redundancia

Además de la tasa de compresión, sobre los archivos se puede calcular el rendimiento y la redundancia. El rendimiento es una medida de la eficiencia del código y la redundancia es su contraparte con el desperdicio. Se calculan mediante la siguiente fórmula:

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

Fórmula 2: Rendimiento.

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

Fórmula 3: Redundancia.

Siendo H(S) la entropía de la fuente y L su longitud media.

Canales de Comunicación

Por otro lado, en la segunda parte se trabajó con canales de comunicación. Un canal de comunicación es una representación matemática que permite modelar el proceso de transmisión y recepción de la información. Se utiliza para medir la calidad de un canal y la aparición de ruido en el mismo. Está determinado por un alfabeto de entrada “A” junto con su probabilidad de emisión, es decir, su probabilidad a-priori P(ai) y uno de salida “B”. También, se cuenta con el conjunto de probabilidades condicionales “P(bj/ai)” que representa la probabilidad de recibir el símbolo “bj” cuando se envía el símbolo de entrada “ai”. Con esta información es posible calcular todas las propiedades del canal como medidas de calidad del mismo.

Propiedades de los Canales

Con la probabilidad a-priori y la matriz de probabilidades condicionales es posible calcular la probabilidad del suceso simultáneo. En este contexto, es la probabilidad de que se dé simultáneamente la emisión del símbolo “ai” y la recepción del símbolo “bj”. Esta se calcula con la siguiente fórmula:

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

Fórmula 4: Probabilidad del suceso simultáneo.

Se puede observar como la primera forma parte de probabilidades que son dato mientras que la segunda de probabilidades calculadas, es decir, que pueden arrastrar error.

Con las probabilidades del suceso simultáneo es posible calcular las probabilidades de los símbolos de salida “bj”. Esta es la probabilidad de observar el símbolo “bj” sin conocer el símbolo emitido. Se calcula con la siguiente fórmula:

$$P(bj) = \sum_{i=1}^{r} P(ai)\,P(bj/ai)$$

Fórmula 5: Probabilidad del símbolo de salida.

Una vez obtenida la probabilidad de cada símbolo se salida es posible calcular la probabilidad a-posteriori P(ai/bj). Esta es la probabilidad de que se haya emitido el símbolo “ai“ sabiendo que se recibió el símbolo “bj”. Se calcula mediante la siguiente fórmula:

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

Fórmula 6: Probabilidad a-posteriori.

La entropía a priori es la entropía de la fuente de entrada fijando cada uno de los elementos del conjunto de salida. Esta se realiza a partir de la probabilidad de los símbolos de la fuente aplicada a la definición de entropía.

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

Fórmula 7: Entropía a-priori.

Al igual que para la entropía a priori, se le aplica a la probabilidad a posteriori la definición de entropía. La interpretación de estas dos entropías (priori y posteriori) puede hacerse basándose en el primer teorema de Shannon. Por un lado, la entropía a priori resulta ser el número medio de binits que se necesitan para poder representar algún símbolo de una fuente con probabilidad a priori, mientras que la probabilidad a posteriori resulta ser el número medio de binits que se necesitan para poder representar algún símbolo de una fuente con una probabilidad a posteriori.

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

Fórmula 8: Entropía a-posteriori.

A raíz de la entropía a posteriori comentada previamente, surge el cálculo de la equivocación. Esta otorga una medida respecto de la pérdida de información sobre “A” causada por el canal. Existen dos formas de calcularla adjuntas continuación:

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

Fórmula 9: Equivocación.

Realizados los cálculos de la entropía a priori y de la equivocación, se puede calcular la información mutua que brinda el canal. Esta es una medida de lo bien o mal que está siendo utilizado un canal.

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

Fórmula 10a: Información mutua primera forma.

$$I(A,B) = \sum_{A,B} P(ai,bj)\ log(\frac{P(ai/bj)}{P(ai)})$$

Fórmula 10b: Información mutua segunda forma.

$$I(A,B) = \sum_{A,B} P(ai,bj)\ log(\frac{P(ai,bj)}{P(ai)P(bj)})$$

Fórmula 10c: Información mutua tercera forma.

Además, la información mutua cumple con las siguientes propiedades:

- I(A,B) >= 0. No se puede perder información al observar la salida. Cuando I(A,B), los símbolos de entrada y salida son estadísticamente independientes.
- Es simétrica respecto a las variables “ai” y “bj”. Entonces I(A,B) = I(B,A).
- Es posible calcular la entropía afin utilizando: H(A,B) = H(A) + H(B) - I(A,B)

A partir de la probabilidad a posteriori puede calcularse la entropía afín. Esta mide la incertidumbre del suceso simultáneo (ai, bj). La probabilidad de este suceso se la representa como P(ai,bj) de modo que la entropía afín valdrá:

$$H(A,B) = \sum_{A,B} P(ai,bj)\ log(\frac{1}{P(ai,bj)})$$

Fórmula 11a: Entropía afín.

Otra forma de calcular la entropía afín es a partir de las siguientes fórmulas:

$$H(A,B) = H(B) + H(A/B)$$

Fórmula 11b: Entropía afín.

$$H(A,B) = H(A) + H(B/A)$$

Fórmula 11c: Entropía afín.

Donde:

- H(B)= Nro. mínimo de preguntas binarias en promedio para determinar la salida.
- H(A/B) =Nro. mínimo de preguntas binarias en promedio para determinar la entrada conocida la salida. Se lo denomina **ruido**.
- H(A) = Nro mínimo de preguntas binarias en promedio para determinar la entrada.
- H(B/A)=Nro. mínimo de preguntas binarias en promedio para determinar la salida conocida la entrada Se lo denomina **pérdida**.

Desarrollo

Compresión de Archivo

Para la primera parte se debió generar dos archivos comprimidos del mismo archivo original, utilizando el algoritmo de Huffman y el de Shannon-Fano. Como se mencionó, estos métodos generan la tabla de codificación dinámicamente. Por lo tanto, se requirió recorrer el archivo original para determinar las probabilidades de aparición de cada palabra. Mediante esta distribución de probabilidad se pudo calcular la tabla de codificación para ambos algoritmos, utilizando los métodos explicados en la introducción. Luego, esta tabla se debió almacenar en el archivo comprimido. Para esto, primero se almacena la cantidad de palabras en 4 bytes y luego cada par de palabra-código. Las palabras se almacenan como bytes terminados con un byte lleno de ceros. Seguido se almacena un byte con la cantidad de bits del código y el código. Esta forma permite un almacenamiento con el mínimo desperdicio posible ya que se realiza bit a bit. Una vez que se almacenó la tabla, el archivo se carga reemplazando las palabras originales por su representación en código. Luego, para descomprimir se recupera la tabla del archivo y se utiliza para reemplazar los códigos por las palabras originales. Con los pesos de los archivos comprimidos se puede calcular la tasa de compresión mediante la fórmula 1:

Huffman	**Shannon-Fano**
73.57%	73.81%

Tabla 1: Tasas de compresión de los archivos comprimidos.

Se puede observar cómo la tasa de compresión del algoritmo de Huffman es mayor. Esto se debe a que el algoritmo de Shannon-Fano genera códigos subóptimos. Esto se condice con los resultados de rendimiento y redundancia de los archivos.

	Original	**Huffman**	**Shannon-Fano**
Rendimiento	33.88%	99.72%	98.53%
Redundancia	66.12%	0.28%	1.47%

Tabla 2: Rendimiento y redundancia para los tres archivos.

Se puede observar cómo el archivo original contiene mucha redundancia. Esto es lo que permite comprimirlo. Además, se puede observar como la redundancia de Huffman es menor que la de Shannon-Fano por la razón mencionada anteriormente.

Canales de Comunicación

Para cada uno de los canales se calculó sus probabilidades de suceso simultáneo utilizando la fórmula 4 para luego calcular las probabilidades de los símbolos de salida utilizando la fórmula 5. Estos resultados se utilizaron para calcular las probabilidades

a-posteriori utilizando la fórmula 6. Con toda esta información fue posible calcular la equivocación mediante dos formas distintas como se puede ver en la fórmula 9, obteniendo de ambas maneras los mismos resultados. Luego se utilizó la fórmula 10 para calcular la información mutua de tres formas distintas. Al igual que con la equivocación, los tres valores resultaron iguales. Por último, se calculó la entropía afín utilizando la fórmula 11. En los tres casos se obtuvieron los mismos resultados.

Canal 1

Los resultados de las entropías a-posteriori son:

	B1	**B2**	**B3**
H(A/bj)	2,2018313899852	2,18475015906035	2,07852864937392

Tabla 3: Entropías a-posteriori del canal 1.

Y los resultados de las propiedades son:

H(A)	**H(B)**	**H(A/B)**	**I(A, B)**	**I(B, A)**	**H(A, B)**	**H(A, B)***
2,1709505	1,5830689	2.1539300	0,0170205	0,0170205	3,7369989	3,7369989

Tabla 4: Propiedades del canal 1.

El resultado de H(A, B)* representa el valor de la entropía afín calculado en base a las propiedades de la información mutua.

En primer lugar, al comparar los valores de la entropía a-priori con las entropías a-posteriori y entendiendo la entropía a-priori cómo la incertidumbre sobre la entrada enviada, se puede observar que la recepción de los símbolos B1 y B2 aumenta la incertidumbre sobre la entrada, mientras que disminuye para la recepción del símbolo B3.

Además, se puede observar que los valores de la entropía a-priori y de la equivocación son muy similares. Esto se refleja en valor bajo para la información mutua, lo que indica que el canal presenta una pérdida de información considerable.

Canal 2

Los resultados de las entropías a-posteriori son:

	B1	**B2**	**B3**	**B4**
H(A/bj)	1.924522883	1.93391091	1.9853085	1.82029336

Tabla 5: Entropías a-posterior del canal 2.

Y los resultados de las propiedades del canal 2 son:

H(A)	H(B)	H(A/B)	I(A,B)	I(B, A)	H(A,B)	H(A, B)*
1.9483888	1.9925638	1.9154523	0.0329365	0.0329365	3.9080161	3.9080161

Tabla 6: Propiedades del canal 2.

El resultado de H(A, B)* representa el valor de la entropía afín calculado en base a las propiedades de la información mutua.

Lo primero que se puede comparar es la entropía a-priori con las entropías a-posteriori. Entendiéndose a la entropía a-priori como la incertidumbre sobre la entrada enviada se puede observar como la recepción del símbolo B3 aumenta la incertidumbre sobre la entrada mientras que disminuye para los símbolos B1, B2 y B4.

Por otro lado, se puede observar que los valores de la entropía a-priori y la equivocación son muy similares. Por lo tanto, el canal posee una perdida de información considerable. Esto se ve reflejado en el valor bajo que toma la información mutua. Esto significa que la información que se obtiene de A gracias a observar B es muy poco considerable.

Canal 3

Los resultados de las entropías a-posteriori son:

	B1	B2	B3	B4
H(A/bj)	2.47627057	2.52594671	2.53938623	2.42904851

Tabla 7: Entropías a-posterior del canal 3.

Y los resultados de las propiedades del canal 3 son:

H(A)	H(B)	H(A/B)	I(A,B)	I(B, A)	H(A,B)	H(A,B)*
2.5272504	1.9946871	2.495965	0.0312854	0.0312854	4.4906521	4.4906521

Tabla 8: Propiedades del canal 3.

El resultado de H(A, B)* representa el valor de la entropía afín calculado en base a las propiedades de la información mutua.

En primer lugar, al comparar los valores de la entropía a-priori con las entropías a-posteriori y entendiendo la entropía a-priori cómo la incertidumbre sobre la entrada enviada, se puede observar que la recepción del símbolo B3 aumenta la incertidumbre sobre la entrada, mientras que disminuye para la recepción de los símbolos B1,B2 y B4.

En segundo lugar, se puede observar que los valores de la entropía a-priori y de la equivocación son muy similares. Esto se refleja en valor bajo para la información mutua, lo que indica que el canal presenta una pérdida de información considerable.

Comparación de los Canales

Si se observan los resultados de los canales comparados:

	H(A)	H(A/B)	I(A,B)
Canal 1	2.17095059445	3.73699893895	0.01702056710
Canal 2	1.948388868	1.915452325	0.032936543
Canal 3	2.52725044	2.495965	0.03128544

Tabla 9: Propiedades de los canales comparadas.

Comparando las propiedades de los canales, se puede observar que el canal 2 presenta la menor equivocación entre los tres. Es decir, posee la menor pérdida de información debido al canal. Además, comparando la información mutua, se puede determinar que el canal 2 es el que obtiene la mayor información sobre la entrada observando la salida.

También se puede observar que cumplen con las propiedades de la información mutua. En ninguno de los tres casos se pierde información al observar las salidas. Además, como se puede observar en las tablas 4, 6 y 8, son simétricas respecto a las variables “ai” y “bj” y se pudo comprobar que la entropía afín está relacionada con la entropía de la entrada, la salida y la información mutua.

Conclusiones

En conclusión, a partir de los resultados obtenidos a lo largo del desarrollo, se observa que tanto Huffman como Shannon-Fano resultaron en una tasa de compresión muy similar. Sin embargo, comparando los resultados podemos notar que Huffman resulta ligeramente más eficiente en medidas de tasa de compresión. También, comparando el rendimiento y la redundancia de los códigos generados por los métodos se puede contemplar que resulta mayor el rendimiento del código generado con el método de Huffman lo cual es consistente con lo mencionado previamente, ya que este es una medida de cuan aprovechada está la información que brinda cada palabra.

Por otra parte, se observa que es posible obtener información acerca de los canales de comunicación conociendo propiedades de los mismos tales como la información mutua y la equivocación. El análisis de estos valores permite conocer si el canal que se utiliza es el indicado para transmitir la información deseada.

Anexo

Canal 1

P(a): Probabilidad a priori		Matriz del canal		
Símbolo	**Probabilidad**	**B1**	**B2**	**B3**
S1	0.2	0.3	0.3	0.4
S2	0.1	0.4	0.4	0.2
S3	0.3	0.3	0.3	0.4
S4	0.3	0.3	0.4	0.3
S5	0.1	0.3	0.4	0.3

Canal 2

Probabilidad a-priori		Matriz del canal			
Símbolo	**Probabilidad**	**B1**	**B2**	**B3**	**B4**
S1	0.25	0.2	0.3	0.2	0.3
S2	0.33	0.3	0.3	0.2	0.2
S3	0.27	0.3	0.2	0.2	0.3
S4	0.15	0.3	0.3	0.3	0.1

Canal 3

Probabilidad a-priori		Matriz del canal			
Símbolo	**Probabilidad**	**B1**	**B2**	**B3**	**B4**
S1	0.15	0.3	0.2	0.3	0.3
S2	0.1	0.3	0.3	0.1	0.3
S3	0.2	0.2	0.3	0.3	0.2
S4	0.25	0.3	0.2	0.2	0.3
S5	0.14	0.3	0.3	0.2	0.3
S6	0.16	0.3	0.3	0.2	0.3